

RECENT ADVANCES in COMPUTER SCIENCE

**Proceedings of the 19th International Conference on Computers
(part of CSCC '15)**

**Zakynthos Island, Greece
July 16-20, 2015**

RECENT ADVANCES in COMPUTER SCIENCE

**Proceedings of the 19th International Conference on Computers
(part of CSCC '15)**

**Zakynthos Island, Greece
July 16-20, 2015**

Copyright © 2015, by the editors

All the copyright of the present book belongs to the editors. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the editors.

All papers of the present volume were peer reviewed by no less than two independent reviewers. Acceptance was granted when both reviewers' recommendations were positive.

Series: Recent Advances in Computer Engineering Series | 32

ISSN: 1790-5109

ISBN: 978-1-61804-320-7

RECENT ADVANCES in COMPUTER SCIENCE

**Proceedings of the 19th International Conference on Computers
(part of CSCC '15)**

**Zakynthos Island, Greece
July 16-20, 2015**

Organizing Committee

Editor:

Prof. Xiaodong Zhuang, Automation & Engineering College, Qingdao University, China

Associate Editors:

Prof. Dr. Eduardo Mario Dias

Prof. Vladimír Vašek

Prof. J. Angela Jennifa Sujana

Prof. Dr. Abdel-Badeeh M. Salem

Prof. José Machado

Prof. Dorota Jelonek

Prof. Nikos Bardis

Prof. V. V. Kozlov

Assoc. Prof. Miroslav Voznak

Dr. N. Rajesh Jesudoss Hynes

Organizing Committee:

Prof. Kleanthis Psarris, The City University of New York, USA (General Chair)

Prof. Pierre Borne, IEEE France Section Chair, IEEE Fellow, Ec Centr de Lille, France (General Chair)

Prof. Panos M. Pardalos, University of Florida, USA (Co-Chair)

Prof. George Vachtsevanos, Georgia Institute of Technology, Atlanta, Georgia, USA (Co-Chair)

Prof. Tadeusz Kaczorek, IEEE Fellow, Warsaw University of Technology, Poland (Co-Chair)

Prof. Nikos Mastorakis, Technical University of Sofia, Bulgaria (Program Chair)

Prof. Branimir Reljin, University of Belgrade, Belgrade, Serbia (International Liaisons)

Prof. Aida Bulucea, University of Craiova, Craiova, Romania (Publicity Chair)

Prof. Valeri Mladenov, Technical University of Sofia, Bulgaria (Publications Chair)

Prof. Imre Rudas, Obuda University, Budapest, Hungary (Tutorials Chair)

Prof. Vladimir Vasek, Tomas Bata University, Zlin, Czech Republic (Special Sessions Chair)

Prof. Anca Croitoru, Al.I. Cuza University, Iasi, Romania (Workshops Chair)

Steering Committee:

Prof. Yuriy S. Shmaliy, IEEE Fellow, Universidad de Guanajuato, Mexico

Prof. Alaa Khamis, IEEE Robotics and Automation Egypt-Chapter Chair, Egypt

Prof. Ioannis Stathopoulos, Technical University of Athens, Greece

Prof. Charalambos Arapatsakos, University of Thrace, Greece

Prof. Fragkiskos Topalis, Technical University of Athens, Greece

Prof. Klimis Ntalianis, Technological Educational Institute of Athens, Greece

Prof. Eduardo Mario Dias, University of Sao Paulo, Brazil

Prof. Miroslav Voznak, VSB-Technical University of Ostrava, Czech Republic

Prof. Abdel-Badeeh M. Salem, Ain Shams University, Cairo, Egypt

Prof. Nikolaos Bardis, M.Inst. of Univ. Educ. (ASEI), Hellenic Army Academy, Athens, Greece

Prof. Antoanela Naaji, Vasile Goldis Western University Arad, Romania

Prof. Elena Zamiatina, Perm State University, Perm Krai, Russia

Prof. Pan Agathoklis, University of Victoria, Canada

Prof. George Tsekouras, M.Inst. of Univ. Educ. (ASEI), Hellenic Naval Academy, Athens, Greece

Prof. Claudio Talarico, Gonzaga University, Spokane, USA

International Scientific Committee:

Prof. Lotfi Zadeh (IEEE Fellow, University of Berkeley, USA)

Prof. Leon Chua (IEEE Fellow, University of Berkeley, USA)

Prof. Michio Sugeno (RIKEN Brain Science Institute (RIKEN BSI), Japan)

Prof. Dimitri Bertsekas (IEEE Fellow, MIT, USA)

Prof. Demetri Terzopoulos (IEEE Fellow, ACM Fellow, UCLA, USA)

Prof. Georgios B. Giannakis (IEEE Fellow, University of Minnesota, USA)
Prof. Abraham Bers (IEEE Fellow, MIT, USA)
Prof. Brian Barsky (IEEE Fellow, University of Berkeley, USA)
Prof. Aggelos Katsaggelos (IEEE Fellow, Northwestern University, USA)
Prof. Josef Sifakis (Turing Award 2007, CNRS/Verimag, France)
Prof. Hisashi Kobayashi (Princeton University, USA)
Prof. Kinshuk (Fellow IEEE, Massey Univ. New Zeland),
Prof. Leonid Kazovsky (Stanford University, USA)
Prof. Narsingh Deo (IEEE Fellow, ACM Fellow, University of Central Florida, USA)
Prof. Kamisetty Rao (Fellow IEEE, Univ. of Texas at Arlington, USA)
Prof. Anastassios Venetsanopoulos (Fellow IEEE, University of Toronto, Canada)
Prof. Steven Collicott (Purdue University, West Lafayette, IN, USA)
Prof. Nikolaos Paragios (Ecole Centrale Paris, France)
Prof. Nikolaos G. Bourbakis (IEEE Fellow, Wright State University, USA)
Prof. Stamatios Kartalopoulos (IEEE Fellow, University of Oklahoma, USA)
Prof. Irwin Sandberg (IEEE Fellow, University of Texas at Austin, USA),
Prof. Michael Sebek (IEEE Fellow, Czech Technical University in Prague, Czech Republic)
Prof. Hashem Akbari (University of California, Berkeley, USA)
Prof. Lei Xu (IEEE Fellow, Chinese University of Hong Kong, Hong Kong)
Prof. Paul E. Dimotakis (California Institute of Technology Pasadena, USA)
Prof. Martin Pelikan (UMSL, USA)
Prof. Patrick Wang (MIT, USA)
Prof. Wasfy B Mikhael (IEEE Fellow, University of Central Florida Orlando, USA)
Prof. Sunil Das (IEEE Fellow, University of Ottawa, Canada)
Prof. Nikolaos D. Katopodes (University of Michigan, USA)
Prof. Bimal K. Bose (Life Fellow of IEEE, University of Tennessee, Knoxville, USA)
Prof. Janusz Kacprzyk (IEEE Fellow, Polish Academy of Sciences, Poland)
Prof. Sidney Burrus (IEEE Fellow, Rice University, USA)
Prof. Biswa N. Datta (IEEE Fellow, Northern Illinois University, USA)
Prof. Mihai Putinar (University of California at Santa Barbara, USA)
Prof. Wlodzislaw Duch (Nicolaus Copernicus University, Poland)
Prof. Michael N. Katehakis (Rutgers, The State University of New Jersey, USA)
Prof. Pan Agathoklis (Univ. of Victoria, Canada)
Dr. Subhas C. Misra (Harvard University, USA)
Prof. Martin van den Toorn (Delft University of Technology, The Netherlands)
Prof. Malcolm J. Crocker (Distinguished University Prof., Auburn University, USA)
Prof. Urszula Ledzewicz, Southern Illinois University, USA.
Prof. Dimitri Kazakos, Dean, (Texas Southern University, USA)
Prof. Ronald Yager (Iona College, USA)
Prof. Athanassios Manikas (Imperial College, London, UK)
Prof. Keith L. Clark (Imperial College, London, UK)
Prof. Argyris Varonides (Univ. of Scranton, USA)
Dr. Michelle Luke (Univ. Berkeley, USA)
Prof. Patrice Brault (Univ. Paris-sud, France)
Prof. Jim Cunningham (Imperial College London, UK)
Prof. Philippe Ben-Abdallah (Ecole Polytechnique de l'Universite de Nantes, France)
Prof. Ichiro Hagiwara, (Tokyo Institute of Technology, Japan)
Prof. Akshai Aggarwal (University of Windsor, Canada)
Prof. Ulrich Albrecht (Auburn University, USA)
Prof. Alexey L Sadovski (IEEE Fellow, Texas A&M University, USA)
Prof. Amedeo Andreotti (University of Naples, Italy)
Prof. Ryszard S. Choras (University of Technology and Life Sciences Bydgoszcz, Poland)
Prof. Remi Leandre (Universite de Bourgogne, Dijon, France)
Prof. Moustapha Diaby (University of Connecticut, USA)

Prof. Brian McCartin (New York University, USA)
Prof. Anastasios Lyrintzis (Purdue University, USA)
Prof. Charles Long (Prof. Emeritus University of Wisconsin, USA)
Prof. Marvin Goldstein (NASA Glenn Research Center, USA)
Prof. Ron Goldman (Rice University, USA)
Prof. Ioannis A. Kakadiaris (University of Houston, USA)
Prof. Richard Tapia (Rice University, USA)
Prof. Milivoje M. Kostic (Northern Illinois University, USA)
Prof. Helmut Jaberg (University of Technology Graz, Austria)
Prof. Ardeshir Anjomani (The University of Texas at Arlington, USA)
Prof. Heinz Ulbrich (Technical University Munich, Germany)
Prof. Reinhard Leithner (Technical University Braunschweig, Germany)
Prof. M. Ehsani (Texas A&M University, USA)
Prof. Sesh Commuri (University of Oklahoma, USA)
Prof. Nicolas Galanis (Universite de Sherbrooke, Canada)
Prof. Rui J. P. de Figueiredo (University of California, USA)
Prof. Hiroshi Sakaki (Meisei University, Tokyo, Japan)
Prof. K. D. Klaes, (Head of the EPS Support Science Team in the MET Division at EUMETSAT, France)
Prof. Emira Maljevic (Technical University of Belgrade, Serbia)
Prof. Kazuhiko Tsuda (University of Tsukuba, Tokyo, Japan)
Prof. Nobuoki Mano (Meisei University, Tokyo, Japan)
Prof. Nobuo Nakajima (The University of Electro-Communications, Tokyo, Japan)
Prof. P. Vanderstraeten (Brussels Institute for Environmental Management, Belgium)
Prof. Annaliese Bischoff (University of Massachusetts, Amherst, USA)
Prof. Fumiaki Imado (Shinshu University, Japan)
Prof. Sotirios G. Ziavras (New Jersey Institute of Technology, USA)
Prof. Marc A. Rosen (University of Ontario Institute of Technology, Canada)
Prof. Thomas M. Gattton (National University, San Diego, USA)
Prof. Leonardo Pagnotta (University of Calabria, Italy)
Prof. Yan Wu (Georgia Southern University, USA)
Prof. Daniel N. Riahi (University of Texas-Pan American, USA)
Prof. Alexander Grebennikov (Autonomous University of Puebla, Mexico)
Prof. Bennie F. L. Ward (Baylor University, TX, USA)
Prof. Guennadi A. Kouzaev (Norwegian University of Science and Technology, Norway)
Prof. Geoff Skinner (The University of Newcastle, Australia)
Prof. Hamido Fujita (Iwate Prefectural University(IPU), Japan)
Prof. Francesco Muzi (University of L'Aquila, Italy)
Prof. Claudio Rossi (University of Siena, Italy)
Prof. Sergey B. Leonov (Joint Institute for High Temperature Russian Academy of Science, Russia)
Prof. Lili He (San Jose State University, USA)
Prof. M. Nasseh Tabrizi (East Carolina University, USA)
Prof. Alaa Eldin Fahmy (University Of Calgary, Canada)
Prof. Gh. Pascovici (University of Koeln, Germany)
Prof. Pier Paolo Delsanto (Politecnico of Torino, Italy)
Prof. Radu Munteanu (Rector of the Technical University of Cluj-Napoca, Romania)
Prof. Ioan Dumitrache (Politehnica University of Bucharest, Romania)
Prof. Miquel Salgot (University of Barcelona, Spain)
Prof. Amaury A. Caballero (Florida International University, USA)
Prof. Maria I. Garcia-Planas (Universitat Politecnica de Catalunya, Spain)
Prof. Petar Popivanov (Bulgarian Academy of Sciences, Bulgaria)
Prof. Alexander Gegov (University of Portsmouth, UK)
Prof. Lin Feng (Nanyang Technological University, Singapore)
Prof. Colin Fyfe (University of the West of Scotland, UK)
Prof. Zhaohui Luo (Univ of London, UK)

Prof. Wolfgang Wenzel (Institute for Nanotechnology, Germany)
Prof. Weilian Su (Naval Postgraduate School, USA)
Prof. Phillip G. Bradford (The University of Alabama, USA)
Prof. Hamid Abachi (Monash University, Australia)
Prof. Josef Boercsoek (Universitat Kassel, Germany)
Prof. Eyad H. Abed (University of Maryland, Maryland, USA)
Prof. Andrzej Ordys (Kingston University, UK)
Prof. T Bott (The University of Birmingham, UK)
Prof. T.-W. Lee (Arizona State University, AZ, USA)
Prof. Le Yi Wang (Wayne State University, Detroit, USA)
Prof. Oleksander Markovskyy (National Technical University of Ukraine, Ukraine)
Prof. Suresh P. Sethi (University of Texas at Dallas, USA)
Prof. Hartmut Hillmer (University of Kassel, Germany)
Prof. Bram Van Putten (Wageningen University, The Netherlands)
Prof. Alexander Iomin (Technion - Israel Institute of Technology, Israel)
Prof. Roberto San Jose (Technical University of Madrid, Spain)
Prof. Minvydas Ragulskis (Kaunas University of Technology, Lithuania)
Prof. Arun Kulkarni (The University of Texas at Tyler, USA)
Prof. Joydeep Mitra (New Mexico State University, USA)
Prof. Vincenzo Niola (University of Naples Federico II, Italy)
Prof. S. Y. Chen, (Zhejiang University of Technology, China and University of Hamburg, Germany)
Prof. Duc Nguyen (Old Dominion University, Norfolk, USA)
Prof. Tuan Pham (James Cook University, Townsville, Australia)
Prof. Jiri Klima (Technical Faculty of CZU in Prague, Czech Republic)
Prof. Rossella Cancelliere (University of Torino, Italy)
Prof. Wladyslaw Mielczarski (Technical University of Lodz, Poland)
Prof. Ibrahim Hassan (Concordia University, Montreal, Quebec, Canada)
Prof. Erich Schmidt (Vienna University of Technology, Austria)
Prof. James F. Frenzel (University of Idaho, USA)
Prof. Vilem Srovnal, (Technical University of Ostrava, Czech Republic)
Prof. J. M. Giron-Sierra (Universidad Complutense de Madrid, Spain)
Prof. Rudolf Freund (Vienna University of Technology, Austria)
Prof. Alessandro Genco (University of Palermo, Italy)
Prof. Martin Lopez Morales (Technical University of Monterey, Mexico)
Prof. Ralph W. Oberste-Vorth (Marshall University, USA)
Prof. Photios Anninos, Democritus University of Thrace, Greece

Additional Reviewers

Bazil Taha Ahmed	Universidad Autonoma de Madrid, Spain
James Vance	The University of Virginia's College at Wise, VA, USA
Sorinel Oprisan	College of Charleston, CA, USA
M. Javed Khan	Tuskegee University, AL, USA
Jon Burley	Michigan State University, MI, USA
Xiang Bai	Huazhong University of Science and Technology, China
Hessam Ghasemnejad	Kingston University London, UK
Angel F. Tenorio	Universidad Pablo de Olavide, Spain
Yamagishi Hiromitsu	Ehime University, Japan
Imre Rudas	Obuda University, Budapest, Hungary
Takuya Yamano	Kanagawa University, Japan
Abelha Antonio	Universidade do Minho, Portugal
Andrey Dmitriev	Russian Academy of Sciences, Russia
Valeri Mladenov	Technical University of Sofia, Bulgaria
Francesco Zirilli	Sapienza Universita di Roma, Italy
Ole Christian Boe	Norwegian Military Academy, Norway
Masaji Tanaka	Okayama University of Science, Japan
Jose Flores	The University of South Dakota, SD, USA
Kazuhiko Natori	Toho University, Japan
Matthias Buyle	Artesis Hogeschool Antwerpen, Belgium
Frederic Kuznik	National Institute of Applied Sciences, Lyon, France
Minhui Yan	Shanghai Maritime University, China
Eleazar Jimenez Serrano	Kyushu University, Japan
Konstantin Volkov	Kingston University London, UK
Miguel Carriegos	Universidad de Leon, Spain
Zhong-Jie Han	Tianjin University, China
Francesco Rotondo	Polytechnic of Bari University, Italy
George Barreto	Pontificia Universidad Javeriana, Colombia
Moran Wang	Tsinghua University, China
Alejandro Fuentes-Penna	Universidad Autónoma del Estado de Hidalgo, Mexico
Shinji Osada	Gifu University School of Medicine, Japan
Kei Eguchi	Fukuoka Institute of Technology, Japan
Philippe Dondon	Institut polytechnique de Bordeaux, France
Dmitrijs Serdjuks	Riga Technical University, Latvia
Deolinda Rasteiro	Coimbra Institute of Engineering, Portugal
Stavros Ponis	National Technical University of Athens, Greece
Tetsuya Shimamura	Saitama University, Japan
João Bastos	Instituto Superior de Engenharia do Porto, Portugal
Genqi Xu	Tianjin University, China
Santoso Wibowo	CQ University, Australia
Tetsuya Yoshida	Hokkaido University, Japan
José Carlos Metrôlho	Instituto Politecnico de Castelo Branco, Portugal

Table of Contents

Plenary Lecture 1: Error Estimation in the Decoupling of Ill-Defined and/or Perturbed Nonlinear Processes	19
<i>Pierre Borne</i>	
Plenary Lecture 2: Applications of Linear Algebra in Signal Processing, Wireless Communications and Bioinformatics	21
<i>Erchin Serpedin</i>	
Plenary Lecture 3: Reliability Life Cycle Management for Engineered Systems	22
<i>George Vachtsevanos</i>	
Plenary Lecture 4: Augmented Reality: The Emerging Trend in Education	24
<i>Minjuan Wang</i>	
Plenary Lecture 5: Application of Multivariate Empirical Mode Decomposition in EEG Signals for Subject Independent Affective States Classification	26
<i>Konstantinos N. Plataniotis</i>	
Plenary Lecture 6: State of the Art and Recent Progress in Uncertainty Quantification for Electronic Systems (i.e. Variation-Aware or Stochastic Simulation)	28
<i>Luca Daniel</i>	
The Evolution of Customer Relationship Management System	29
<i>Dorota Jelonek</i>	
Big Data Analytics of Social Media	34
<i>Peter Wlodarczak, Jeffrey Soar, Mustafa Ally</i>	
Design of Methodology for Connecting Enterprise Architect with Development Solutions and Necessary Application Framework	40
<i>J. Sedivy, R. Borkovec, P. Coufal</i>	
New Challenges in Smart Campus Applications	44
<i>Attila Adamkó, Tamás Kádek, Lajos Kollár, Márk Kósa, János Pánovics</i>	
Machine-Learning - An Overview of Optimization Techniques	51
<i>Pedro Oliveira, Filipe Portela, Manuel Filipe Santos, António Abelha, José Machado</i>	
The Personalized Recommendation Technology for Online Courses with Combinational Algorithm	57
<i>Minjuan Wang, Jun Xiao, Bingqian Jiang, Junli Li</i>	

Theoretical Analysis and Experimental Evaluation of Bandwidth Amplification Attacks to Legitimate Websites	63
<i>Dimitrios P. Iracleous, Kristofer E. Bourro, Nikolaos Doukas</i>	
Digital Image Segmentation Inspired by Carrier Immigration in Physical P-N Junction	68
<i>Xiaodong Zhuang, Nikos E. Mastorakis</i>	
Designing Engaging Mobile Learning for K-12 Classrooms	75
<i>Minjuan Wang, Melissa Calderwood, Yong Chen, Junli Li</i>	
How to Break Down the Security of an Efficient Modular Exponentiation Algorithm	81
<i>David Tinoco Varela</i>	
Objective Stimulus Features for Predicting Human Judgments of Visual Pattern Goodness: An Empirical Comparison	86
<i>Godfried T. Toussaint</i>	
Extending Cloud Computing and Learning for Mobility	92
<i>Phil Robisch, Rebecca J. Kirsininkas, Minjuan Wang</i>	
Research on the Analytics Model Design of Online Learning Behavior	97
<i>Jun Xiao, Minjuan Wang, Lamei Wang, Bingqian Jiang</i>	
Student Anxiety Awareness through a Bio-Feedback Device as a Significant Support to Educational Activities	103
<i>Hippokratīs Apostolidis, Thrasylvoulos Tsiatsos, Minjuan Wang</i>	
Bio-Inspired Algorithms for Attack of Block Ciphers	108
<i>T. Mekhaznia, A. Zidani</i>	
Influence of Mesh Quality and Density on Numerical Calculation of Heat Exchanger with Undulation in Herringbone Pattern	115
<i>Václav Dvořák, Jan Novosád</i>	
Sampling Time Dependency of Chaotic Ueda Oscillator as the Generator of Random Numbers for Heuristic	121
<i>Roman Senkerik, Michal Pluhacek, Zuzana Kominkova Oplatkova</i>	
The Application of Business Intelligence Systems in the Support of Decision Processes in the International Enterprises	127
<i>Leszek Ziara</i>	
The Concept of a Model of the Separation of the User Interface Layer from the Database Layer in B2B System	131
<i>M. Łobaziewicz</i>	

Implementing the Green Transport Strategy Using Balanced Scorecard and Analytic Network Process	139
<i>D. Staš, R. Lenort, P. Wicher, D. Holman</i>	
Information Technologies in Logistics Services. Case Study	145
<i>Izabela Krawczyk-Sokołowska, Katarzyna Łukasik</i>	
The Development of E-Business Services in Poland	151
<i>Elzbieta Wyslocka, Renata Biadacz</i>	
Fourth Dimension of Spatial Description in Business Processes	157
<i>Cezary Stępniaak</i>	
Design and Implementation of the Korean Style Plug-In Using the Wordpress	163
<i>Jeongseok Ji, Jaesic Kim, Youngwan Kim, Sungjin Jung, Chaehyun Lee, Dongsu Kim, Yonggoon Kim, Miyoung Bae, Yangwon Lim, Hankyu Lim</i>	
A Comparison of Open-Source CMS - Focused on the CMS Market Place in Korea -	167
<i>Yangwon Lim, Youseck Yang, Hyeonpyo Hong, Geunwoo Ahn, Jeongwoo Lee, Eunju Park, Jihyeon Hwang, Yonggoon Kim, Hankyu Lim</i>	
Support for Reports and Forms Printing in wxWidgets GUI Toolkit	170
<i>Michal Bližňák, Tomáš Dulík, Roman Jašek</i>	
Automation of Modern Marketing Tools	177
<i>Dagmara Bubel</i>	
System for Professionals – Monitoring Employers’ Demands for Key Competences in Wielkopolska	184
<i>M. Szafranski, M. Goliński</i>	
Data Assimilation Method Coupled with the Numerical Simulation of the Ocean Dynamics	192
<i>Konstantin P. Belyaev, Andrey A. Kuleshov, Clemente A. S. Tanajura, Natalia P. Tuchkova</i>	
Implementation of a Kinetically-Based Algorithm for Porous Medium Flow Simulation on Hybrid Supercomputers	197
<i>Andrew A. Kuleshov, Natalia G. Churbanova, Anastasiya A. Lyupa, Marina A. Trapeznikova</i>	
Efficient Distribution Conversion Algorithm in Low Power TRNGs for Embedded Security Applications	201
<i>Blerim Rexha, Dren Imeraj, Ehat Qerimi, Arbnor Halili</i>	
Successive Elimination Algorithm for Truncated Gray-Coded Bitplane Matching Based Motion Estimation	206
<i>Ilseung Kim, Jechang Jeong</i>	

Improving Programming Courses Using Aptitude Testing and Learning Styles <i>Eva Milková, Karel Petránek</i>	211
Interactive Teaching Tools for Visualizing Geometrical 3D Objects Using Pseudo Holographic Images <i>M. Ciobanu, A. Ploscar, I. Dascal, I. Virag, A. Naaji</i>	215
Implications of Domain-Driven Design in Complex Software Value Estimation and Maintenance Using DSL Platform <i>Nikola Vlahovic</i>	219
Comparative Advantages of Software Industry in Developing Countries: Study of Structure, Market Strategies and Software Development Approaches in Croatian Software Companies <i>Nikola Vlahovic, Ljubica Milanovic Glavan, Anja Frankovic</i>	227
The Application of Computer Technology in Optimizing the Conditions of Directional Breaking of Fibrous Collagen Linkages <i>Shalbuev Dm. V., Zharnikova E. V., Radnaeva V. D.</i>	234
Building Rich User Profile Based on Intentional Perspective <i>Sara Alaoui, Younès El Bouzekri El Idrissi, Rachida Ajhoun</i>	238
Augmented Reality in Radiofrequency Ablation of the Liver Tumours <i>Lucio Tommaso De Paolis, Francesco Ricciardi, Cosimo Luigi Manes</i>	243
Management of Intangible Assets within Health Care Industry. A Comparative Study between Sweden and Poland <i>Dorota Jelonek, Amra Halilovic</i>	249
Automatic Acquisition, Processing and Analysis of Data System, Using the AHP Multi-Criteria Method <i>Sorin Borza, Carmen Simion</i>	254
Port Operation – Increase of Automated Systems, Decline of Workforce Jobs? <i>Aureo E. P. Figueiredo, Ricardo de D. Carvalhal, Sérgio Hoeflich, Letícia Figueiredo, Sergio L. Pereira, Eduardo M. Dias</i>	259
Implementation of Track and Trace System for Medication in the Largest Hospital Complex in Brazil <i>Elcio B. Da Silva, Maria L. R. P. Dias, Eduardo M. Dias, Sergio L. Pereira</i>	267
A Tabu Search Using Guide Trees-Based Neighborhood for the Multiple Sequence Alignment Problem <i>Tahar Mehenni</i>	275

Computational Automation in Modern Personalized Medicine - AirPROM Project Perspective	281
<i>Michal Kierzyńska, Marcin Adamski, Andreas Fritz, Dmitriy Galka, Ian Jones, Dieter Maier, Andrew Wells</i>	
Intrusion Detection System in Area of Interest Using a Background Subtraction-Based Tracking Algorithm	286
<i>Hanbyul Chae, Kicheon Hong</i>	
Exploratory Social Network Analysis with Pajek: Case Study on Student Group Performance	292
<i>Lionel Khalil, Marie Khair, Tina Daaboul, Marie-Joelle El Hajje</i>	
Warden 3: Security Event Exchange Redesign	298
<i>Pavel Kachá, Michal Kostěněc, Andrea Kropáčová</i>	
Ransomware	304
<i>Jan Kolouch, Andrea Kropáčová</i>	
Face Depth Estimation Using Differential Evolution and Iterative Soft Thresholding Algorithm	308
<i>K. Punnam Chandar, T. Satya Savithri</i>	
Multi-Lane Traffic Flow Models Accounting for Different Lane Changing Motivations	314
<i>M. N. Smirnova, D. A. Pestov, A. I. Bogdanova, N. N. Smirnov, A. B. Kiselev, V. F. Nikitin, V. V. Tyurenkova</i>	
Potential of Pervasive Computing through Embedded Systems and Internet Technologies: Research of Customer Perspective	320
<i>Nikola Vlahović, Jovana Zoroja, Vesna Bosilj Vukšić</i>	
Exploiting the Interpretability of Fuzzy Rule-Based Classifiers for Analyzing Hyperspectral Remotely Sensed Data	327
<i>Dimitris G. Stavrakoudis, Stelios K. Mylonas, Charalampos A. Topaloglou, John B. Theocharis, Paris A. Mastorocostas</i>	
Depth Estimation from Single Face Image Using Modified Differential Evolution	335
<i>K. Punnam Chandar, T. Satya Savithri</i>	
The Use of Virtual Laboratory Works at the Teaching of Natural Sciences Subjects	340
<i>Yevgeniya A. Daineko, Madina T. Ipalakova, Viktor G. Dmitriyev, Andrey D. Giyenko, Nazgul K. Rakhimzhanova</i>	
Benefits of Knowledge Engineering for E-Learning Systems	343
<i>Abedl-Badeeh M. Salem, Thakaa Z. Mohamad</i>	

Decision Support System for Predicting Football Game Result	348
<i>João Gomes, Filipe Portela, Manuel Filipe Santos</i>	
Prediction of Potential Organ Donation after Irreversible Brain Damage	354
<i>Luís Torres, Filipe Portela, Manuel Filipe Santos, António Abelha, José Neves, José Machado</i>	
Semantify Educational Resources Using SKOS and Learning Object Ontologies	360
<i>Georgia D. Solomou, Dimitrios A. Koutsomitropoulos, Aikaterini K. Kalou, Sotirios D. Botsios</i>	
Big Data Solutions to Support Intelligent Systems and Applications	366
<i>Luciana Lima, Filipe Portela, Manuel Filipe Santos, António Abelha, José Machado</i>	
Proposed Runtime Decision Making Framework for Autonomic Software Systems	371
<i>Sandeep Kumar Chauhan, Arun Sharma, P. S. Grover</i>	
Modular System for Gathering and Classification of SIP Attacks	376
<i>J. Safarik, M. Voznak, J. Slachta, L. Macura, F. Rezac, J. Rozhon</i>	
A Four-State Markov Chain and its Application in Packet Loss Modelling for Speech Quality Estimation of IP Telephony	382
<i>J. Rozhon, F. Rezac, M. Voznak, J. Safarik, J. Slachta, L. Macura</i>	
Experimental Analysis of the Effects of Turbulent Jets in Shallow Water Bodies	388
<i>Robles L. Isidro, Palacio P. Arturo, Rodríguez V. Alejandro</i>	
Design of M2M Service Capability for Access to Location Information	394
<i>Ivaylo I. Atanasov, Evelina N. Pencheva</i>	
Performance Estimation of Non-Comparison Based Sorting Algorithms Under Different Platforms and Environments	400
<i>Mentor Hamiti, Diellza Nagavci</i>	
Reduced Permissions Schema for Malware Detection in Android Smartphones	406
<i>Ahmed H. Mostafa, Marwa M. A. Elfattah, Aliaa A. A. Youssif</i>	
Fetal Heart Rate Estimation from Phonocardiograms Using an EMD Based Method	414
<i>Dragos Daniel Taralunga, Mihaela Ungureanu, Bogdan Hurezeanu, Rodica Strungaru</i>	
Soft-Error-Rate Adaptive Intervals for Low Overhead Checkpoint	418
<i>Wentao Jia, Chunyuan Zhang, Kun Jiang</i>	
A Novel Technique to Detect and Recognize Faces in Multi-View Videos	427
<i>Steven Lawrence Fernandes, G. Josemin Bala</i>	
A Comparative Study to Recognize Surgically Altered Images	434
<i>Steven Lawrence Fernandes, G. Josemin Bala</i>	

Extraction of Blood Vessels and Optic Disc Segmentation for Retinal Disease Classification	440
<i>Jestin V. K.</i>	
Fuzzy Logic Based Performance Analysis of Various Multiplier Architectures	445
<i>Vardhana M.</i>	
Renovation CoReVDO® Methodology of Collaborative Requirements Validation in Distributed Organizations	449
<i>Sourour Maalem</i>	
Interactive Image Search for Mobile Devices	457
<i>Komal V. Aher, Sanjay B. Waykar</i>	
Benefits of New Laboratory Tools in Research and Education	463
<i>Gabriela Gladiola Andruseac, Mădălina Poștaru, Corina Cheptea, Anca-Irina Galaction</i>	
A Probabilistic Clustering-Based Adaptive Histogram Thresholding Method for Fast Segmentation of Color Images	469
<i>Abolfazl Mirkazemy, S. Enayatolah Alavi, Gholamreza Akbarizadeh</i>	
Big Data Analytics in Prevention, Preparedness, Response and Recovery in Crisis and Disaster Management	476
<i>Dontas Emmanouil, Doukas Nikolaos</i>	
Identifying Peer-to-Peer Traffic Based on Traffic Characteristics	483
<i>S. R. Patil, Suraj Sanjay Dangat</i>	
Influence of IT on Micro Enterprises to Pursue Strategic Growth	488
<i>Satya Shah, Syed Hassan</i>	
Binarization and Recognition of Characters from Historical Degraded Documents	497
<i>Bency Jacob, S. B. Waykar</i>	
Adaptive Analysis of Characteristic Nodes Using Prediction Method in DTN	502
<i>A. Yoon-Hyung Dho, Kang-Whan Lee</i>	
Cyber Diversity for Security of Digital Substations under Uncertainties: Assurance and Assessment	507
<i>E. Brezhniev, V. Kharchenko, J. Vain, A. Boyarchuk</i>	
Green Computing within the Context of Educational and Research Projects	513
<i>Vyacheslav Kharchenko, Oleg Illiashenko, Chris Phillips, Jüri Vain</i>	
Evolution of Software Quality Models: Usability, Security and Greenness Issues	519
<i>Oleksandr Gordieiev, Vyacheslav Kharchenko, Mario Fusani</i>	

Simulation on Friction Welding Of MgAZ31 / AA 6061 T6 Joints <i>N. Rajesh Jesudoss Hynes, P. Shenbaga Velu</i>	524
Authors Index	529

Plenary Lecture 1

Error Estimation in the Decoupling of Ill-Defined and/or Perturbed Nonlinear Processes



Professor Pierre Borne (IEEE Fellow)

Co-authors Amira Gharbi, Mohamed Benrejeb

Centre de Recherche en Informatique Signal et Automatique de Lille, CRISTAL

Ecole Centrale de Lille

France

E-mail: pierre.borne@ec-lille.fr

Abstract: This lecture deals with the definition of the attractors characterizing the precision of decoupling control laws for a nonlinear process in presence of uncertainties and/or bounded perturbations. This approach is based on the use of aggregation techniques and the definition of a comparison system of the controlled process.

Brief Biography of the Speaker: Pierre BORNE received the Master degree of Physics in 1967 and the Master of Electrical Engineering, the Master of Mechanics and the Master of Applied Mathematics in 1968. The same year he obtained the Diploma of "Ingenieur IDN" (French "Grande Ecole"). He obtained the PhD in Automatic Control of the University of Lille in 1970 and the DSc in physics of the same University in 1976. Dr BORNE is author or co-author of about 200 Publications and book chapters and of about 300 communications in international conferences. He is author of 18 books in Automatic Control, co-author of an english-french, french-english « Systems and Control » dictionary and co-editor of the "Concise Encyclopedia of Modelling and Simulation" published with Pergamon Press. He is Editor of two book series in French and co-editor of a book series in English. He has been invited speaker for 40 plenary lectures or tutorials in International Conferences. He has been supervisor of 76 PhD Thesis and member of the committee for about 300 doctoral thesis . He has participated to the editorial board of 20 International Journals including the IEEE, SMC Transactions, and of the Concise Subject Encyclopedia . Dr BORNE has organized 15 international conferences and symposia, among them the 12th and the 17 th IMACS World Congresses in 1988 and 2005, the IEEE/SMC Conferences of 1993 (Le Touquet – France) and of 2002 (Hammamet - Tunisia) , the CESA IMACS/IEEE-SMC multiconferences of 1996 (Lille – France) , of 1998 (Hammamet – Tunisia) , of 2003 (Lille-France) and of 2006 (Beijing, China) and the 12th IFAC LSS symposium (Lille France, 2010) He was chairman or co-chairman of the IPCs of 34 international conferences (IEEE, IMACS, IFAC) and member of the IPCs of more than 200 international conferences. He was the editor of many volumes and CDROMs of proceedings of conferences. Dr BORNE has participated to the creation and development of two groups of research and two doctoral formations (in Casablanca, Morocco and in Tunis, Tunisia). twenty of his previous PhD students are now full Professors (in France, Morocco, Tunisia, and Poland). In the IEEE/SMC Society Dr BORNE has been AdCom member (1991-1993 ; 1996-1998), Vice President for membership

(1992-1993) and Vice President for conferences and meetings (1994-1995, 1998-1999). He has been associate editor of the IEEE Transactions on Systems Man and Cybernetics (1992-2001). Founder of the SMC Technical committee « Mathematical Modelling » he has been president of this committee from 1993 to 1997 and has been president of the « System area » SMC committee from 1997 to 2000. He has been President of the SMC Society in 2000 and 2001, President of the SMC-nomination committee in 2002 and 2003 and President of the SMC-Awards and Fellows committee in 2004 and 2005. He is member of the Advisory Board of the "IEEE Systems Journal" . Dr. Borne received in 1994, 1998 and 2002 Outstanding Awards from the IEEE/SMC Society and has been nominated IEEE Fellow the first of January 1996. He received the Norbert Wiener Award from IEEE/SMC in 1998, the Third Millennium Medal of IEEE in 2000 and the IEEE/SMC Joseph G. Wohl Outstanding Career Award in 2003. He has been vice president of the "IEEE France Section" (2002-2010) and is president of this section since 2011. He has been appointed in 2007 representative of the Division 10 of IEEE for the Region 8 Chapter Coordination sub-committee (2007-2008) He has been member of the IEEE Fellows Committee (2008- 2010) Dr BORNE has been IMACS Vice President (1988-1994). He has been co-chairman of the IMACS Technical Committee on "Robotics and Control Systems" from 1988 to 2005 and in August 1997 he has been nominated Honorary Member of the IMACS Board of Directors. He is since 2008 vice-president of the IFAC technical committee on Large Scale Systems. Dr BORNE is Professor "de Classe Exceptionnelle" at the "Ecole Centrale de Lille" where he has been Head of Research from 1982 to 2005 and Head of the Automatic Control Department from 1982 to 2009. His activities concern automatic control and robust control including implementation of soft computing techniques and applications to large scale and manufacturing systems. He was the principal investigator of many contracts of research with industry and army (for more than three millions €) Dr BORNE is "Commandeur dans l'Ordre des Palmes Académiques" since 2007. He obtained in 1994 the french " Kulman Prize". Since 1996, he is Fellow of the Russian Academy of Non-Linear Sciences and Permanent Guest Professor of the Tianjin University (China). In July 1997, he has been nominated at the "Tunisian National Order of Merit in Education" by the Republic of Tunisia. In June 1999 he has been nominated « Professor Honoris Causa » of the National Institute of Electronics and Mathematics of Moscow (Russia) and Doctor Honoris Causa of the same Institute in October 1999. In 2006 he has been nominated Doctor Honoris Causa of the University of Waterloo (Canada) and in 2007 Doctor Honoris Causa of the Polytechnic University of Bucharest (Romania). He is "Honorary Member of the Senate" of the AGORA University of Romania since May 2008 He has been Vice President of the SEE (French Society of Electrical and Electronics Engineers) from 2000 to 2006 in charge of the technical committees. He his the director of publication of the SEE electronic Journal e-STA and chair the publication committee of the REE Dr BORNE has been Member of the CNU (French National Council of Universities, in charge of nominations and promotions of French Professors and Associate Professors) 1976-1979, 1992-1999, 2004-2007 He has been Director of the French Group of Research (GDR) of the CNRS in Automatic Control from 2002 to 2005 and of a "plan pluriformations" from 2006 to 2009. Dr BORNE has been member of the Multidisciplinary Assessment Committee of the "Canada Foundation for Innovation" in 2004 and 2009. He has been referee for the nominations of 24 professors in USA and Singapore. He is listed in the "Who is Who in the World" since 1999.

Plenary Lecture 2

Applications of Linear Algebra in Signal Processing, Wireless Communications and Bioinformatics



Professor Erchin Serpedin

Department of Electrical and Computer Engineering
Texas A&M University
USA

E-mail: serpedin@ece.tamu.edu

Abstract: In this talk, we will review some of the most important applications of linear algebra in signal processing, wireless communications and bioinformatics, and then outline some of the major open problems which might benefit by the usage of linear algebra concepts and tools.

Brief Biography of the Speaker: Dr. Erchin Serpedin is currently a professor in the Department of Electrical and Computer Engineering at Texas A&M University in College Station. He is the author of 2 research monographs, 1 textbook, 9 book chapters, 105 journal papers and 175 conference papers. Dr. Serpedin serves currently as associate editor for the Physical Communications Journal (Elsevier) and EURASIP Journal on Advances in Signal Processing, and as Editor-in-Chief of the journal EURASIP Journal on Bioinformatics and Systems Biology edited by Springer. He is an IEEE Fellow and his research interests include signal processing, biomedical engineering, bioinformatics, and machine learning.

Plenary Lecture 3

Reliability Life Cycle Management for Engineered Systems



Professor George Vachtsevanos

Professor Emeritus

Georgia Institute of Technology

USA

E-mail: george.vachtsevanos@ece.gatech.edu

Abstract: Engineered systems are becoming more complex and by necessity more unreliable resulting in detrimental events for the system itself and its operator. There is evidence to support the contention that industrial and manufacturing processes, transportation and aerospace systems, among many others, are subjected to severe stresses, external and internal, that contribute to increased cost, operator workload, frequent and catastrophic mishaps that require the development and application of new and innovative technologies to improve system reliability, safety, availability and maintainability. These requirements are not true only for strictly engineered systems but are often discussed in business and finance, biological systems and social networks. We introduce in this talk a systematic and verifiable methodology to improve system reliability, reduce operating costs and optimize system design or maintenance practices. The enabling technologies build upon modeling tools to represent critical system functions, a prognostic strategy to predict the long-term behavior of systems under stress, reliability analysis methods exploiting concepts of probabilistic design and an optimization algorithm to arrive at optimum system design for improved reliability. We demonstrate the efficacy of the approach with examples from the engineering domain.

Brief Biography of the Speaker: Dr. George Vachtsevanos is currently serving as Professor Emeritus at the Georgia Institute of Technology. He served as Professor of Electrical and Computer Engineering at the Georgia Institute of Technology from 1984 until September, 2007. Dr Vachtsevanos directs at Georgia Tech the Intelligent Control Systems laboratory where faculty and students began research in diagnostics in 1985 with a series of projects in collaboration with Boeing Aerospace Company funded by NASA and aimed at the development of fuzzy logic based algorithms for fault diagnosis and control of major space station subsystems. His work in Unmanned Aerial Vehicles dates back to 1994 with major projects funded by the U.S. Army and DARPA. He has served as the Co-PI for DARPA's Software Enabled Control program over the past six years and directed the development and flight testing of novel fault-tolerant control algorithms for Unmanned Aerial Vehicles. He has represented Georgia Tech at DARPA's HURT program where multiple UAVs performed surveillance, reconnaissance and tracking missions in an urban environment. Under AFOSR sponsorship, the Impact/Georgia Team is developing a biologically-inspired micro aerial vehicle. His research work has been supported over the years by ONR, NSWC, the MURI Integrated Diagnostic

program at Georgia Tech, the U.S. Army's Advanced Diagnostic program, General Dynamics, General Motors Corporation, the Academic Consortium for Aging Aircraft program, the U.S. Air Force Space Command, Bell Helicopter, Fairchild Controls, among others. He has published over 300 technical papers and is the recipient of the 2002-2003 Georgia Tech School of ECE Distinguished Professor Award and the 2003-2004 Georgia Institute of Technology Outstanding Interdisciplinary Activities Award. He is the lead author of a book on Intelligent Fault Diagnosis and Prognosis for Engineering Systems published by Wiley in 2006.

Plenary Lecture 4

Augmented Reality: The Emerging Trend in Education



Professor Minjuan Wang

San Diego State University

USA

E-mail: mwang@mail.sdsu.edu

Abstract: Augmented Reality (AR) is the layering of virtual information over the real, 3-D world to produce a blended reality. AR has been considered a significant tool in education for many years. It adds new layers of interactivity, context, and information for learners which can deepen and enrich the learning experience. The combination of real and virtual allows the student to engage in learning about a topic from multiple perspectives and data sources at levels that are not always available in traditional classroom settings and interactions.

As the usage of mobile devices in formal settings continues to rise, so does the opportunity to harness the power of augmented reality (AR) to enhance teaching and learning. Many educators have experimented with AR, but has it proven to improve what students grasp and retain? Is AR just another fun way to engage students, with little transformation of learning? This plenary speaking will introduce augmented reality as an emerging trend in education, provide an overview of its current development, explore examples of curriculum integration, and also suggest approaches for success.

Brief Biography of the Speaker: Dr. Minjuan Wang (Professor of San Diego State University; Distinguished Research Professor of Shanghai International Studies University)

Homepage: <http://www.tinyurl.com/minjuan>

Minjuan is Professor of Learning, Design, and Technology at San Diego State University (California, USA), and distinguished professor of Shanghai International Studies University (Shanghai, China). She was recently selected as the “Oriental Scholar” by the Municipal Educational Committee of Shanghai). In addition, she and her American colleagues obtained a four-year 1.3 million grant to study environment protection (including the Golden monkeys) in Fanjingshan, Guizhou province.

Minjuan’s work has been highly interdisciplinary, covering the field of education, technology, computer science, geography, and communication. In her 14 years at SDSU, she teaches Designing and Developing Learning for the Global Audience, Mobile Learning Development, Technologies for Course Delivery, and Methods of Inquiry. Her research specialties focus on online learning, mobile learning, Cloud Learning, and intelligent learning (as part of the Intelligent Camps initiative launched by British Telecom). Minjuan is the Editor-in-Chief of a newly established journal-- EAI Transactions on Future Intelligent Educational Environments. She also serves on the editorial boards for four indexed journals: Open Education Research,

International Journal on E-Learning (IJEL), the Open Education Journal, and Journal of Information Technology Application in Education.

As a winner of several research awards, Minjuan is recognized as one of the high impact authors in blended and mobile learning. She has more than 80 peer-reviewed articles published in indexed journals, such as Educational Technology Research and Development, IEEE Transactions on Education, and British Journal of Educational Technology. She was a keynote and invited speaker to 11 international conferences. In addition, she is also an accomplished creative writer and an amateur flamenco dancer. Her recent Novel--Walking in this Beautiful World—has inspired many young people around the world.

Plenary Lecture 5

Application of Multivariate Empirical Mode Decomposition in EEG Signals for Subject Independent Affective States Classification



Prof. Konstantinos N. Plataniotis

Department of Electrical and Computer Engineering
University of Toronto
CANADA

E-mail: kostas@ece.utoronto.ca

Abstract: Physiological signals, EEG in particular, are inherently noisy and non-linear in nature which are challenging to work with using conventional linear signal processing methods. In this paper, we are adopting a new signal processing method, Multivariate Empirical Mode Decomposition, as a preprocessing method to reconstruct EEG signals according to its instantaneous frequencies. To test its effectiveness, we applied this signal reconstruction technique to analyze EEG signals for a 2-dimensional affect states classification application. To evaluate the proposed EEG signal processing system, a three-class classification experiment was carried out on the “Emobrain” dataset from eINTERFACE'06 with K-nearest neighbors (KNN) and Linear Discriminate Analysis (LDA) as classifiers. A leave-one-subject out cross validation process was used and an averaged correct classification rate of 90.77% was achieved. Another main contribution of this paper was inspired by the growth of non-medical grade EEG headsets and its potential in advanced human-computer interface design. However, to reduce cost and invasiveness, consumer grade EEG headsets have far less number of electrodes. In this paper, we used emotion recognition as a case study, and applied Genetic Algorithm to systematically select the critical channels (or sensor locations) for this application. The results of this study will shed light on the sensor configuration challenges faced by most consumer-grade EEG headset design projects.

Brief Biography of the Speaker: Konstantinos N. (Kostas) Plataniotis received his B. Eng. degree in Computer Engineering from University of Patras, Greece and his M.S. and Ph.D. degrees in Electrical Engineering from Florida Institute of Technology Melbourne, Florida. He was with the Computer Science Department at Ryerson University, Ontario, Canada from July 1997 to June 1999. Dr. Plataniotis is currently a Professor with The Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto in Toronto, Ontario, Canada, where he directs the Multimedia Laboratory. He is a founding member and the inaugural Director – Research of the Identity, Privacy and Security Institute, IPSI, (www.ipsi.utoronto.ca). Kostas was the Director (January 2010- June 2012) of the Knowledge Media Design Institute, KMDI, (www.kmdi.utoronto.ca) at the University of Toronto.

Dr. Plataniotis was the Guest Editor for the March 2005 IEEE Signal Processing Magazine special issue on “Surveillance Networks and Services”, and the Guest Editor for the EURASIP Applied

Signal Processing Journal's special issue on "Advanced Signal Processing & Pattern Recognition Methods for Biometrics". He is a member of the IEEE Periodicals Review and Advisory Committee (2011-2013); he has served as a member of the 2008 IEEE Educational Activities Board; he chaired of the IEEE EAB Continuing Professional Education Committee, and he served as the 2008 representative of the Computational Intelligence Society to the IEEE Biometrics Council. Dr. Plataniotis chaired the 2009 Examination Committee for the IEEE Certified Biometrics Professional (CBP) Program (www.ieeebiometricscertification.org) and he served on the Nominations Committee for the IEEE Council on Biometrics. He was a member of the Steering Committee for the IEEE Transaction on Mobile Computing, an Associate Editor for the IEEE Signal Processing Letters as well as the IEEE Transactions on Neural Networks and Adaptive Systems and he has served as the Editor-in-Chief for the IEEE Signal Processing Letters from January 1, 2009 to December 31, 2011. Dr. Plataniotis chaired the IEEE Toronto Signal Processing and Applications Toronto Chapter from 2000 to 2002, he was the 2004-05 Chair of the IEEE Toronto Section and a member of the 2006 as well as 2007 IEEE Admissions & Advancement Committees. He served as the Technical Program Committee Co-Chair for the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013) and he is the Vice President – Membership for the IEEE Signal Processing Society (2014-2016). Dr. Plataniotis is a Fellow of IEEE, Fellow of the Engineering Institute of Canada, a registered professional engineer in the province of Ontario, and a member of the Technical Chamber of Greece.

The recipient of numerous grants and research contracts as the principal investigator, he speaks internationally and writes extensively in his field and he has been a consultant to a number of companies. He has served as lecturer in 12 short courses to industry and continuing education programs; he is a contributor to seventeen books, the co-author of "Color Image Processing and Applications", Springer Verlag, 2000, (ISBN-3-540-66953-1) and "WLAN Positioning Systems: Principles & applications in Location-based Services", Cambridge University Press, 2012 (ISBN 978-0-521-9185-2), "Multi-linear Subspace Learning: Reduction of multi-dimensional data", CRC Press, 2013, (ISBN: 978-14398557243). He is the co-editor of "Color Imaging: Methods and Applications", CRC Press, September 2006, (ISBN 084939774X) and the Guest Editor of the IEEE/Wiley Press volume on "Biometrics: Theory, Methods and Applications" published in October 2009 (ISBN: 9780470247822). Dr. Plataniotis has published more than 400 papers in refereed journals and conference proceedings. In 2005 he became the recipient of the IEEE Canada Engineering Educator Award for "contributions to engineering education and inspirational guidance of graduate students". Dr. Plataniotis is the joint recipient of the "2006 IEEE Trans. on Neural Networks Outstanding Paper Award" for the published in 2003 "Face recognition using kernel direct discriminant analysis algorithms", IEEE Trans. on Neural Networks, Vol. 14, No 1, 2003.

Plenary Lecture 6

State of the Art and Recent Progress in Uncertainty Quantification for Electronic Systems (i.e. Variation-Aware or Stochastic Simulation)



Professor Luca Daniel

Electrical Engin. & Computer Science
Massachusetts Institute of Technology (MIT)
Cambridge, MA, USA
E-mail: luca@mit.edu

Abstract: On-chip and off chip fabrication process variations have become a major concern in today's electronic systems design since they can significantly degrade systems' performance. Existing commercial circuit and MEMS simulators mostly rely on the well known Monte Carlo algorithm in order to predict and quantify such performance degradation. However during the last decade a large variety of more sophisticated and efficient alternative approaches have been proposed to accelerate such critical task. This talk will first review the state of the art of most modern uncertainty quantification techniques including intrusive and sampling-based ones. It will be shown in particular how parameterized model order reduction, and low-rank tensor based representations can be used to accelerate most uncertainty quantification tools and to handle the curse of dimensionality. Examples will be presented including amplifiers, mixers, voltage controlled oscillators with tunable micro-electro-mechanical capacitors and phase locked loops.

Brief Biography of the Speaker: Luca Daniel is an Associate Professor in the Electrical Engineering and Computer Science Department of the Massachusetts Institute of Technology (MIT). Prof. Daniel received the Ph.D. degree in Electrical Engineering from the University of California, Berkeley, in 2003. In 1998, he was with HP Research Labs, Palo Alto. In 2001, he was with Cadence Berkeley Labs.

Dr. Daniel research interests include development of integral equation solvers for very large complex systems, stochastic field solvers for large number of uncertainties, and automatic generation of parameterized stable compact models for linear and nonlinear dynamical systems. Applications of interest include simulation, modeling and optimization for mixed-signal/RF/mm-wave circuits, power electronics, MEMs, nanotechnologies, materials, MRI, and the human cardiovascular system.

Prof. Daniel has received the 1999 IEEE Trans. on Power Electronics best paper award; the 2003 best PhD thesis awards from both the Electrical Engineering and the Applied Math departments at UC Berkeley; the 2003 ACM Outstanding Ph.D. Dissertation Award in Electronic Design Automation; 5 best paper awards in international conferences, 8 additional nominations for best paper award; the 2009 IBM Corporation Faculty Award; and the 2010 IEEE Early Career Award in Electronic Design Automation.

The Evolution of Customer Relationship Management System

Dorota Jelonek

Abstract – Paper presents the evolution of Customer Relationship Management Systems from the classical solutions, through e-CRM systems, to social CRM. The aim of this article is to demonstrate that social CRM systems are an effective support in managing the relationships with customers, especially in the areas of customer information management and customer communication.

Keywords— CRM, e-CRM, s-CRM, customer communication, customer information management

I. INTRODUCTION

Strategies oriented on the customer and strengthening customer relationships allow modern enterprises to get a competitive advantage on the market and make a bigger profit. This means that companies should develop their skills in terms of identification of customer needs and expectations and then provide customers with more and more benefits and satisfaction resulting therefrom. The role of customer is showed in the concepts of co-creating value with customers discussed by Prahalad i Ramaswamy [1], P. Kotler and K. Keller [2], P.F. Drucker [3] and in the concepts of innovation co-creation e.g. open innovations Chesbrough [4], Jelonek [5], collective intelligence Glenn [6] or crowdsourcing Howe [7].

Creation and evolution of customer relationships are a condition for cooperation.

Customer Relationship Management (CRM) is both a business strategy and information system, which will increase the effectiveness of the implementation of the strategy. Enterprises leverage the latest information technology achievements in the development of their long-term relationships with customers [8], [9].

The model and functionality of the CRM system changes with the development of ICT, especially with the development of the internet. Model e-CRM can be as activities to manage customer relationships by using the internet, web browsers or other electronic touch points. The popularity of social media has caused the CRM systems to evolve towards social CRM systems (s-CRM). s-CRM uses social media to develop and sustain interaction between customers and company.

The purpose of this paper is to demonstrate that social CRM systems are an effective support in managing the relationships

D. Jelonek is with the Faculty of Management, Czestochowa University of Technology, Czestochowa, Poland (+48343250846; e-mail: jelonek@zim.pcz.pl).

with customers, especially in the areas of customer information management and customer communication.

The following paragraphs present the essence of CRM systems, e-CRM and s-CRM models as well as the role of s-CRM in customer information management and customer communication.

II. THE ESSENCE AND FUNCTIONS OF CRM SYSTEM

Customer Relationship Management in the literature of the subject is considered as a strategy [10], process [11], philosophy [12], skill [13] or system [14]. Thus, the essence of CRM was well-defined by Greenberg [15]: „CRM (...) is not only technology. It is a strategy and/or a set of business processes. A methodology. It is all of the above or whichever you choose”.

CRM may be defined as the cross-functional integration of processes, people, operations, and marketing capabilities that is enabled through information, technology and applications [16].

Due to the CRM functions it can be divided into three basic types [17]:

- operational CRM,
- analytical CRM,
- collaborative CRM.

Figure 1 shows the model of CRM system that includes operational, analytical and collaborative modules.

Operational CRM, often referred to as front-office CRM, covers most areas of customer - company contact. CRM applications collect, process and store data about customers, so that later this data can be used in analytical CRM [18].

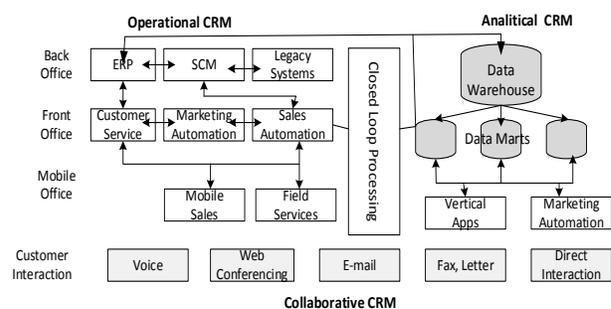


Fig. 1. The model of CRM

Analytical CRM, also known as back-office CRM, uses data from operational CRM and any other sources of data such as transaction systems or enterprise environment. Most of today's CRM vendors develop their own analytical CRM

modules or collaborates with producers of specialized information processing systems of the Business Intelligence type.

Collaborative CRM, also called the interactive CRM, are applications, that support various forms of contact with customers, especially by using modern technologies of electronic communication. The usage of ICT supports the work of employees who contact directly with customers, allowing for partial automation of these contacts.

The usage of internet in business and changes in the virtual environment made it necessary to modify CRM system. Network CRM, referred to as e-CRM, uses internet technologies, and like traditional CRM it implements processes of acquiring, storing and processing information about e-customers as well as sharing them with managers.

The possibilities of e-CRM, in terms of broadly understood customer service, may include [19]: building lasting relationships with e-customers, increasing the level of e-customer satisfaction, boosting sales, identification of those e-customers who generate highest or lowest revenues, minimizing costs of e-customer services, benefits from retail, decreasing costs of customer management, acquiring new e-customers, more efficient customer service resulting from personalization of service, providing e-customers with full information, creation of possibilities to choose, understanding e-customer needs, effective marketing communication with e-customers and quicker access to new markets.

III. MODEL OF S-CRM SYSTEM

The evolution of Web 2.0 and social media have significantly changed the customer relationship management model towards social CRM.

Social Media can be defined as a group of internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user generated content [20]. They are the technical enabler for an online-based exchange of digital contents and operationalize the principles and practices of Web 2.0. As the front-end to the user they represent something tangible compared to the intangible concept of Web 2.0. [21]. Customers want to talk about their consumption experience, new ideas, however they have various preferences on where, what and how to communicate. s-CRM system should allow for a full dialogue with customers using the communication channels of their preference. Moreover, web-user integration and participation becomes critical to establish trust and commitment in buyer-seller relationships.

Social CRM system may be defined from various perspectives. Social CRM is a philosophy and a business strategy, supported by a technology platform, business rules, processes and social characteristics, designed to engage the customer in a collaborative conversation in order to provide mutually beneficial value in a reliable and transparent business. It's the company response to the customer's property on the conversation [22].

According to Mohan [23], a social CRM system combines the "Web 2.0 features and social networking with current

CRM system." However, Social CRM is not just a set of technologies, but rather a company strategy, specific to boost customer engagement and building strong relationships with them. Askool and Nakata [24] describe SCRMM to be even a new paradigm for creating high value relationships.

s-CRM definitions point out that s-CRM is more than an extension of traditional CRM by means of new communication channels and about a new mode of managing relationships in a public environment that builds on and integrating the principles and practices of Web 2.

New dimension of social CRM add to the traditional aspects of customer relationship management was presented in Figure 2.

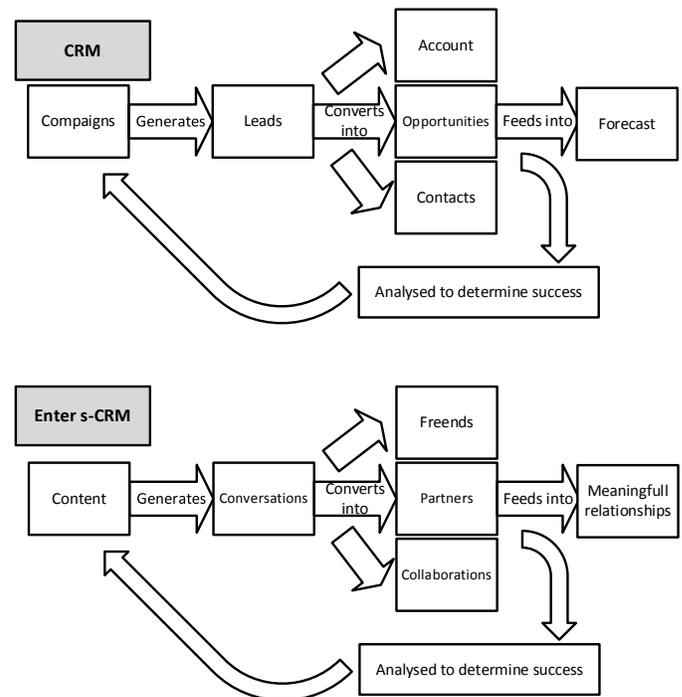


Figure 2. CRM v. s-CRM approach [25]

There were showed three important processes:

- Converting content into conversation.
- Extending conversation into collaborative experience.
- Transforming experience into meaningful relationship.

The main social media are clearly Facebook, Twitter, YouTube, LinkedIn and now Instagram and Pinterest [26]. However, the nature of the Internet through Web 2.0, is that almost all websites are becoming "social"; that is that they allow interaction [27]. Customers can post comments, ratings, reviews, and share all these across their own social networks. Thus, sites like Amazon, TripAdvisor, Urbanspoon, Yelp, the entire Google network and many other peer-to-peer websites such as blogs, micro blogs, wikis, podcasts, photo sharing, video sharing and social bookmarking can be viewed as social media technologies and thus as potential CRM tools [27], [28].

How an organization can use social channels to support s-CRM processes was presented on table 1.

Table 1. How an organization can use social channels to support s-CRM processes

Social channels	CRM and entity process		
	marketing	sales	service
Blog	-blog focused on building reputation written by a senior executive -focus on leadership	deals with members of a community, for acquisition of products and services	-capture of comments in the executive blog regarding claims or requests by customers, and to act accordingly
Internal Wiki	platform to share market knowledge that has been collected from conversations with customers	shared presentations on sales and common knowledge about new sales leads	creation of a knowledge base of customer service procedures
Video Sites/ YouTube	viral advertising propagated only on-line encouraging word-of-mouth references	point of contact to create sales opportunity in another channel	publication of educational videos on how to use certain product, extending the user manual on-line
Micro Blogging/ Twitter	-messages to announce special offers and discounts -spreading of viral marketing campaigns, integration with channels like YouTube	-launching of exclusive product offers for Twitter followers, as a way of looking for new sales opportunities -focus on the current follower base	-response to support inquires and product complaints, monitored by an exclusive team -focus on all digital customers
Personal Social Networks/ (Facebook)	-spreading of advertising campaigns within the communities of clients	-launch of new product and benefit campaigns for community members only -focus on the current follower base	-resolution of enquires among community members -opinion gathering regarding products and services

Source: [29]

The comparison of key differences between CRM and s-CRM in terms of their functions and features was presented in the Table 2.

Table 2. CRM v. social CRM. Features and Functions

CRM Features/Functions	s-CRM Features/Functions
Definition: CRM is a philosophy & a business strategy, supported by a system and a technology, designed to improve human interactions in a business environment	Definition: s-CRM (CRM 2.0) is a philosophy & a business strategy, supported by a system and a technology, designed to engage the customer in a collaborative interaction that provides mutually beneficial value in a trusted & transparent business environment
Tactical and operational: Customer strategy is part of corporate strategy	Strategic: Customer strategy IS corporate strategy
Relationship between the company and the customer was seen as enterprise managing	Relationship between the company and the customer are seen as a collaborative effort. And yet, the

customer - parent to child to a large extent	company must still be an enterprise in all other aspects
Focus on Company Customer Relationship	Focus on all iterations of the relationships (among company, business partners, customers) and specifically focus on identifying, engaging and enabling the "influential" nodes
The company seeks to lead and shape customer opinions about products, services, and the company-customer relationship.	The customer is seen as a partner from the beginning in the development and improvement of products, services, and the company-customer relationship
Business focus on products and services that satisfy customers	Business focus on environments & experiences that engage customer
Customer facing features - sales, marketing & support.	Customer facing both features and the people who's in charge of developing and delivering those features
Marketing focused on processes that sent improved, targeted, highly specific corporate messages to customer	Marketing focused on building relationship with customer - engaging customer in activity and discussion, observing and re-directing conversations and activities among customers
Intellectual Property protected with all legal might available	Intellectual property created and owned together with the customer, partner, supplier, problem solver
Insights and effectiveness were optimally achieved by the single view of the customer (data) across all channels by those who needed to know. Based on "complete" customer record and data integration	Insights are a considerably more dynamic issue and are based on 1) customer data 2) customer personal profiles on the web and the social characteristics associated with them 3) customer participation in the activity acquisition of those insights
Resided in a customer-focused business ecosystem	Resides in a customer ecosystem
Technology focused around operational aspects of sales, marketing, support	Technology focused on both the operational and social aspects of the interaction
Tools are associated with automating functions	Integrates social media tools into apps/services: blogs, wikis, podcasts, social networking tools, content sharing tools, user communities, tools are associated with communicating
Utilitarian, functional, operational	Style and design also matter
Mostly uni-directional	Always bi-directional
Based on a toolset (software)	Based on a strategy (corporate culture)

Source: [30]

IV. THE ROLE OF S-CRM IN CUSTOMER INFORMATION MANAGEMENT

In order to meet the needs of customers effectively companies must maintain a level of engagement with customers, but they must also be able to acquire and manage information on their customers [31]. Information management includes the following activities: information capture; information integration; information access, and information use. Social media and virtual communities collect a lot of data,

that can be captured. Data can indicate market trends, customer preferences, customer satisfaction, customer influence and value, and competitor information [28], [26].

Information integration requires the assimilation of customer information from all touch points, from different data sources, not just social media, to create a coherent picture of the customers, develop a single view of the customer and collect information about their interaction with the organization.

s-CRM system supports decision-making process only when delivered analysis are based on information resources of all data-collecting systems that are used in enterprise. The format and way of sharing information is also very important. It should be adjusted, so that a company can use information as quickly as possible.

CRM relies on the historical data based on previous buying cycles and experiences of clients. In traditional CRM approach data is logged by third party, usually with some time gap after the event has occurred. Whereas, s-CRM approach incorporates real-time data for real time information (ability to capture unforeseen sales opportunities) and data is generated directly by customers (better reliability of information).

With every engagement with customers on social media, more data is created [32].

In general, Social Media offers five different resources for s-CRM [33]:

1. The content of a posting (Posting Body) can be analysed for key words, opinions, topics, etc.
2. The meta data of postings (Posting Envelope) can provide details about authors, topics, sources, etc.
3. Provided data in profiles (Profile Body) contains information about emails, phone numbers, hobbies, interests, etc.
4. Meta data of profiles (Profile Envelope) contains information about friends, activities, other profiles, etc.
5. Interconnections between postings and profiles (Links) can provide insight into a person role, influence or relations.

Profile body and envelope, Posting envelope and Links are often available as structured data that may be integrated with CRM data by existing functionalities of CRM systems. Posting bodies and implicit links are unstructured data that needs to be transformed by data or text-mining (TM) before they can be integrated with CRM systems [33]. Data listed above is being used by the analytical module of s-CRM.

s-CRM support process of information management ensuring high quality customer information. Various customer information sources are well integrated and the customer information provided by system is useful. Moreover, the customer scoring and segmentation information are supported by CRM system.

V. THE ROLE OF S-CRM IN CUSTOMER COMMUNICATION PROCESS

Companies strengthen customer relationships by adjusting the communication system to clients expectations. Customer are increasingly using communication possibilities of Web 2.0

and Social Media, therefore they expect companies to do the same. Instead of pursuing a traditional one-way push communication, organizations are expected to foster a two-way interaction. At present, consumer expectations are likely not to be fulfilled and there is a perception gap on intentions to use Social Media [21].

CRM has traditionally consisted of one-way communication between company and the customer. s-CRM system assumes continuous exchange of experiences, not only between company and client, but also between individual clients.

Changes in the communication model were presented on the Figure 3. Instead of one-way communication companies should conduct dialog with clients and collaborate with them.

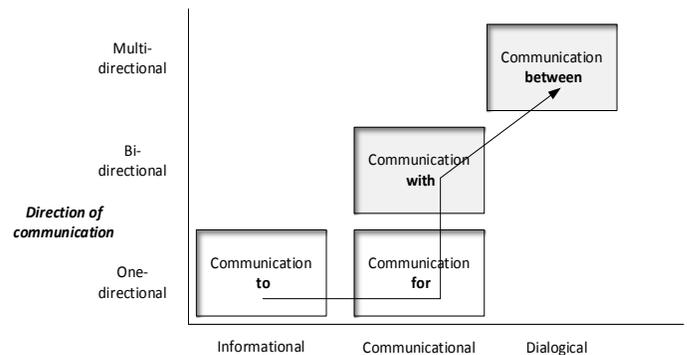


Fig. 3. Web 2.0 communication matrix [34]

Communication “to” the target groups is characterized by pushing persuasive messages of promotion, selling propositions towards a mass market (e.g. online ads).

Communication “for” is a target group approach planned persuasive messages with augmented offerings (e.g. a discount due to a membership anniversary).

Communication “with” rests on bi-directional interactions between an organization and its target groups and emphasizes an exchange of planned and spontaneous messages which is product and service focused [21].

Communication “between” emphasizes dialogue including an organization and multiple consumers.

s-CRM enables businesses to interact with customers in real time using variety of social media platforms in a cost effective way. Companies can use blogs, internal Wiki, video sites, micro blogging, personal social networks and effectively conduct their activities connected with sales, marketing, and customer services.

Social media as a platform for communication offers customers the role of promoters, commentators or co-creators. In other words, customers can become the primary actors in the process of creating the proposals.

VI. CONCLUSION

As more and more consumers are active on social media, marketers’ use of these tools is also increasing. 21 % of marketers say that social media has become more important to their company over the past six months [35]. Marketers have

found a customers via Facebook (52%), LinkedIn (43%), Twitter (36%) [35]. Social media technologies as the element of CRM have the capability to bring company and customers closer together through two-way interactions and dialog.

It was showed that s-CRM systems may effectively support customer relations and usage of its potential. It is important to perceive s-CRM system from the perspective of: supporting customer information management and improving dialogue with customers by using new channels of communication. Newly acquired customers can become in the future: the ambassadors of the brand, reviewers, innovators or consultants. Social media creates new behaviors of community members, such as sharing of experiences and emotions. Companies can later use those behaviors in order to build and strengthen relationships with customers.

REFERENCES

- [1] C. K. Prahalad, V. Ramaswamy, *Future of Competition: Co-Creating Unique Value with Customers*, Harvard Business School Press Books, 2003.
- [2] P. Kotler, K. Keller, *Marketing Management (13th Edition)*, Prentice Hall, 2008.
- [3] P. F. Drucker, *Management Challenges for the 21st Century*, New York: HarperBusiness, 1999.
- [4] H. Chesbrough, "The Era of Open Innovation", *MIT Sloan Management Review*, vol. 44/3, pp. 35-41, Spring 2003.
- [5] D. Jelonek, "The Role of the Internet in Open Innovations Models Development", *Business Informatics*, no 1 (23), pp. 38-47, 2012.
- [6] J. C. Glenn, "Collective Intelligence - One of the Next Big Things", *Futura. Finnish Society for Futures Studies*, Helsinki, Finland, vol. 4, 2009.
- [7] J. Howe, "The rise of crowdsourcing", *Wired Magazine*, vol. 14(6), 2006, <http://www.wired.com/wired/archive/14.06/crowds.html> (2015.01.20)
- [8] Nowakowska, J. Nowakowska-Grunt, "Current Status and Perspectives of Implementation of CRM System in Polish Businesses", *CRM. Customer Relationship Management '07. The International Scientific Conference. Lazne Bohdaneč, Czech Republic*, pp.136-144, 2007.
- [9] Mesjasz-Lech, „Wykorzystanie zintegrowanych systemów informatycznych ERP i CRM w przedsiębiorstwach w kontekście logistyki”, *Zeszyty Naukowe Ekonomiczne Problemy Usług Uniwersytet Szczeciński*, vol. 808, pp. 389-398, 2014.
- [10] Gordon, "CRM is a strategy not a tactic", *Ivey Business Journal*, September/October 2001. <http://iveybusinessjournal.com/> (2015.01.07).
- [11] J. Chen, K. Popovich, "Understanding customer relationship management (CRM): People processes and technology". *Business Process Management Journal*, Vol. 9(5), pp. 672 – 688, 2003.
- [12] M. Hasan, "Ensure success of CRM with a change in mindset", *Marketing Management*, vol. 37(8), 2003.
- [13] D. Peppers, M. Rogers, B. Dorf, "Is your company ready for one -to-one marketing?", *Harvard Business Review*, vol. 77(1), 101 – 119, 1999.
- [14] D. Adenbajo, "Classifying and selecting e-CRM applications: An analysis-based proposal", *Management Decision*, Vol. 41(6), pp. 570 – 577, 2003.
- [15] P. Greenberg, *CRM at the speed of light*. Berkeley: McGraw-Hill, 2001.
- [16] Payne, P. Frow, "A strategic framework for customer relationship management", *Journal of Marketing*, vol. 69(4), pp. 167-176, 2005.
- [17] D. Jelonek, "The Role of the CRM System in the Development of the Customer Capital of An Enterprise", *Processes of Capital Supply in Production Enterprises. Joint Work Edited by Helena Kościelniak, Serie Monographs No 1*, pp.103-109, 2006.
- [18] D. Jelonek, A. Chluski, „Możliwości wykorzystania systemów CRM w zakładach opieki zdrowotnej”, [in:] *Technologie informatyczne w administracji publicznej i służbie zdrowia*. Red. J. Goliński, A. Kobyliński, A. Sobczak, Wydawnictwo SGH, Warszawa, pp. 35-47, 2010.
- [19] D. Jelonek, „Zarządzanie relacjami z klientami w wirtualnym otoczeniu organizacji”, *Studia i Prace Kolegium Zarządzania i Finansów. Szkoła Główna Handlowa*, No 136, pp.19-31, 2014.
- [20] Kaplan, M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media", *Business Horizons*, vol. 53(1), pp. 59–68, 2010.
- [21] T. Lehmkühl, "Towards Social CRM - A Model for Deploying Web 2.0 in Customer Relationship Management", Bamberg 2014. [http://verdi.unisg.ch/www/edis.nsf/SysLkpByIdentifier/4288/\\$FILE/dis4288.pdf](http://verdi.unisg.ch/www/edis.nsf/SysLkpByIdentifier/4288/$FILE/dis4288.pdf) (2015-03-01)
- [22] V. Nitu, C. Tileaga, A. Ionescu, "Evolution of CRM in s-CRM", *Economics, Management, and Financial Markets*, Vol. 9(1), pp. 303–310, 2014.
- [23] S. Mohan, E. Choi, D. Min, "Conceptual Modeling of Enterprise Application System Using Social Networking and Web 2.0. Social CRM System", *Convergence and Hybrid Information Technology ICHIT '08. International Conference*, 2008.
- [24] S. Askool, K. Nakata, „A conceptual model for acceptance of social CRM systems based on a scoping study”. *AI & SOCIETY*, vol. 26(3), pp. 205–220, (2010).
- [25] http://crm2.typepad.com/brents_blog/2008/07/social-crm-in-p.html (2015.02.10)
- [26] P. Harrigan M. Miles, "From e-CRM to s-CRM. Critical factors underpinning the social CRM activities of SMEs", *Small Enterprise Research*, vol. 21, issue 1, pp.99-116, 2014.
- [27] M. Chau, J. Xu, "Business intelligence in blogs: understanding consumer interactions and communities". *MIS Quarterly*, vol. 36, pp. 1189-1216, 2012.
- [28] H. Chen, R.H.L. Chiang, V.C. Storey, "Business intelligence and analytics: from big data to big impact", *MIS Quarterly*, vol. 36, pp. 1165-1188, 2012.
- [29] <http://crm2.pbworks.com/> (2015.02.216)
- [30] CRM 2.0 or Social CRM for Finacial Industry, http://www.deloitte.com/assets/Dcom-Croatia/Local%20Assets/Documents/2012/FSINews03.02-Social_CRM.pdf. (2015.02.10)
- [31] V. Hutchinson, P. Quintas, "Do SMEs do knowledge management? Or simply manage what they know?", *International Small Business Journal*, vol. 26, pp. 131-154, 2008.
- [32] T. H. A. Bijmolt, P. S. H. Leeftang, F. Block, M. Eisenbeiss, B.G.S. Hardie, A. Lemmens, P. Saffert, "Analytics for customer engagement", *Journal of Service Research*, vol. 13, pp. 341-356, 2010.
- [33] O. Reinhold, R. Alt, "Social Customer Relationship Management: State of the Art and Learnings from Current Projects", *25th Bled eConference eDependability: Reliable and Trustworthy eStructures, eProcesses, eOperations and eServices for the Future*, Bled, Slovenia, 2012.
- [34] Ballantyne, R. J. Varey, "Introducing a dialogical orientation to the service-dominant logic of marketing". In R. F. Lusch & S. L. Vargo (Eds.), *The Service-Dominant Logic of Marketing: Dialog, Debate, and Directions* (pp. 224–235). Armonk, NY: M.E. Sharpe Inc, 2006.
- [35] HubSpot, 2013 "State of Inbound Marketing Annual Report", <http://offers.hubspot.com/2013-state-of-inbound-marketing> (2014.01.20).

Dorota Jelonek is a professor of Management and currently the Vice Dean of Science at the Faculty of Management at the Czestochowa University of Technology. She started to work at the Faculty of Management in 1994 as an assistant. In 2000 she received a PhD degree. PhD research topic was "Modeling of Information Resources for Enterprise Environment Monitoring System". Dorota Jelonek holds her habilitation in Economic Theory at the Faculty of Management, Information Science and Finances, Wrocław University of Economics in 2011 year. Habilitation study focused on the "Strategic Alignment Between Environment Monitoring and Information Technology in a Company. A Methodological and Empirical Study". She has been the chairman of the team responsible for e-learning implementation in Czestochowa University of Technology since 2012 year. She is the author of 3 books, editor of 6 books. In addition, she is the author or co-author of 130 articles in Polish and foreign journals and book chapters. She participated in over 80 conferences in many research centers. Her scientific and research interests focus on solving problems related to the implementation of management information systems in enterprises, improving management information processes, and computer-assisted monitoring of the business environment.

Prof. Jelonek is a Member of Informing Science Institute (ISI) and a Member of the Board of Scientific Society of Economic Informatics.

Big Data analytics of Social Media

Peter Wlodarczak, Jeffrey Soar, Mustafa Ally

Abstract— Opinions are key influencers of human behavior. Before buying a new car or a camera people, often ask the opinions of friends or acquaintances. In the past years, the Internet has become a major source of information about products and services and for reviews and experience reports. Since the advent of Web 2.0 technologies the Internet has seen an unprecedented amount of opinionated content in forums, blogs and Social Media such as Twitter and Facebook that people increasingly consult before making purchasing decisions or choosing travel destinations. Companies have realized the potential of Social Media data for personalized marketing, for getting user opinions on their products and services, for detecting new trends and business opportunities and for making predictions about market developments. Analyzing Social Media content has become a very active area of research and poses many interesting research questions. It has the potential of changing the way companies do business. However, there are challenges due to the large volumes of data and the velocity at which they are created, and Big Data technologies are required to process them efficiently. There are also challenges in using Social Media data due to the peculiarities of these media such as fake opinions and spam, jargon and slang used in posts or special characters and emoticons that are widely in use. This paper describes the state-of-the-art techniques that have been used in recent research to analyze Social Media content for opinion mining and for making predictions and proposes an approach for Social Media mining based on machine learning techniques.

Keywords— Machine learning, opinion mining, predictive analytics, Social Media

I. INTRODUCTION

SINCE the advent of Web 2.0 technologies, the Web has seen a shift from publisher created to user created content [1]. Web 2.0 and *Social Media* (SM) facilitated the publishing of content by omitting the need to be able to program. Everyone can now post opinions, views, ideas and interests on any topic and they are accessible in real time from anywhere in the world. Facebooks' data volume grows by more than 500 TB every day [2]. On Twitter, more than 500 million Tweets are sent per day by Twitters own account [3]. This resulted in an unprecedented amount of opinionated data globally accessible for anyone from anywhere. Not surprisingly analyzing SM data has become a very active area of research since mining people's opinions can reveal relevant market research information that result in more targeted business

P. Wlodarczak is a research student at the Faculty of Business, Education, Law and Arts, University of South Queensland, Australia (corresponding author phone: 076-488-5774; e-mail: wlodarczak@gmail.com).

J. Soar is a professor at the Faculty of Business, Education, Law and Arts, University of South Queensland, Australia (e-mail: Jeffrey.soar@usq.edu.au).

M. Ally is a lecturer in Information Systems at the School of Management and Enterprise at the University of South Queensland, Australia (e-mail: allym@usq.edu.au).

decisions. SM analysis has also been used for making predictions on the development of financial markets [4], box office sales [5] or disease outbreaks [6] to name a few. To effectively analyze the large volumes of data, *Big Data* techniques have to be applied. First the SM data has to be analyzed for its opinion polarity. *Opinion mining* techniques are often adopted using *Machine Learning* (ML) techniques. ML is a growing area of data analysis. ML schemes are trained using historic data mimicking human learning. Once trained the ML scheme is applied to new, unseen data to make predictions. For instance, an ML algorithm can learn from past customers who switched to a new company to predict which customers are likely to change in the future.

Opinion mining, also called *sentiment analysis*, is a type of *natural language processing* (NLP). It analyses people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes [21]. The emerging research area of opinion mining deals with computational methods in order to find, extract and systematically analyze people's opinions, attitudes and emotions towards certain topics [7]. Its aim is to *classify* documents, SM posts, according to their *sentiment polarity*. The classification can be *binary*, for instance positive or negative user reviews, or *multiclass*, where posts such as Tweets are divided according to mood states such as "excited", "skeptical" or "angry". Opinions can be expressed at the *entity level*, a product as a whole, for instance "the new Tesla is excellent", or at the *aspect level*, for instance "the voice quality of the new iPhone is good but battery life time is short", where individual features of an entity are evaluated. Opinion mining on SM has not only been used in academia, there is a growing interest from the industry to find out what users think of their products or services, to detect trends and find new business opportunities. In the past companies had to conduct surveys to collect and assess customer satisfaction. Using SM, there is no need to issue questionnaires to a sample set of users since all data can be analyzed. This process is also called *SM listening*.

ML is an area of *Artificial Intelligence* (AI). ML techniques detect patterns in data and can adapt when exposed to new data. For instance spam filters often use ML algorithms since they can adapt when new types of spam appear. Opinion mining combined with ML techniques has been used in many domains. A prominent success story was the football finals in Brazil, where Google correctly predicted the winner of 11 out of the 12 final games using ML techniques [13]. Microsoft's Cortana even correctly predicted the winners of all finals [14], however less is known about their predictive model.

ML is a well-studied area, and ML techniques have been

applied to many Big Data analysis problems. However applied to SM there are challenges due to the large volumes and the variety of the data and due to the peculiarities of SM such as slang and jargon used in posts. This paper describes the state-of-the-art Big Data analysis techniques that have been adopted in recent studies to mine opinions in SM and make predictions based on historic SM data. It proposes a four phase approach for collecting and analyzing SM data and to make predictions based on ensemble learning.

II. PREVIOUS WORK

SM analysis has been used in many domains. Sentiment analysis is a growing area of SM mining. Nowadays social media services such as Twitter and Facebook are increasingly used by online users to share and exchange opinions, providing rich resources to understand public opinions [15]. *Social correlation theories* have been proposed for sentiment analysis by some authors [15]. Other studies have used computational approaches for opinion mining. Different opinion mining algorithms have been analyzed and investigated for their effectiveness [7]. Sentence splitting, stemming, part of speech tagger and parsing algorithms were applied. The researchers concluded that extensive text preprocessing and using algorithms that can effectively process noisy content performed best. Machine learning (ML) techniques such as *supervised methods* based on *naïve Bayesian* and *Support Vector Machine* classification as well as *unsupervised methods* using part of speech tagging have been proposed for political opinion mining on SM [16]. ML techniques have also been used for target oriented opinion mining using a *bag-of-words* supervised classifier [17]. The researchers achieved a classification accuracy of 0.69 for classifying Tweets. Other approaches used *Latent Dirichlet Allocation* (LDA) [18]. LDA characterizes every document by a *Dirichlet distribution*. The similarity between documents is then calculated using a distance measure. The authors concluded that the best results were achieved using a *Jaccard index*.

A very active area of research is *predictive analysis* using SM data. Twitter Tweets have been analyzed to make predictions of financial indicators based on public mood states [4]. The authors investigated if there is a *correlation* between certain public moods on Twitter and the development for the Dow Jones Industrial Average (DJIA) using time series analysis. They concluded that certain mood states do correlate with the development of the DJIA. Other studies analyzed whether box-office revenue could be predicted [5]. They concluded that there is a correlation between the number of positive Tweets and box-office revenue. They also found a correlation between the number of Tweets about a movie and the number of spectators. Similar results for stock price and movie box office revenue were obtained by other studies [12] correlating Twitter based time series.

Sentiments can be expressed with emoticons, they have been used for sentiment analysis [8]. Emoticons have been treated

similarly to sentiment words to determine the sentiment polarity of SM posts replacing facial expression in person to person interaction.

An important step in SM analysis is *data pre-processing*. Bitter experience shows that real data is often disappointingly low in quality [9]. Text quality can have a significant impact on the opinion mining process and has been analyzed for several algorithms [7]. Several studies developed improved techniques for purifying SM data from noise and irrelevant content. LDA has been used for relevance filtering [12]. LDA is based on *Latent Semantic Indexing* [11]. It creates a latent description of relevant posts that is used to filter out irrelevant content. This paper builds on previous studies and proposes a methodology described in the next chapter.

III. RESEARCH METHODOLOGY

SM mining encompasses four phases, a *data collection* phase, a *data pre-processing* phase, a *data mining* phase and a *post-processing* phase. The first two phases comprise the *data conditioning* tasks where the data is collected and preprocessed for analysis. In the analysis phase, the data is mined for actionable *patterns* and *correlations* are searched for [11]. In the post-processing phase, the data is often visualized, or reports are generated. In this phase sometimes *predictive analysis* is performed, it is also called the *predictive phase*. It is executed when data is not only mined to understand the underlying structure and detect patterns, but when projections of future events are sought for. Each phase can go through several iterations. Data mining typically goes through many iterations until satisfactory results are achieved. The four phases are described in the next chapters.

A. Data collection

Data of the big SM sites such as Facebook or Google+ can be accessed through *Application Programming Interfaces* (API). The data can thus be accessed programmatically using Java, Python or any other programming or script language. For instance, Facebook has a Graph API that can be used for posting and retrieving data. Twitter has a query API to access historic tweets. Twitter also provides a streaming API to access real-time data. The “firehose” API gives access to 100%, the “gardenhose” API to 10% and the “spritzer” API to 1% of real-time Tweets. Recently Facebook has also added a streaming API to its interfaces. However on many SM sites free access is usually limited. Full access such as Twitters “firehose” API is usually very costly, only the “gardenhose” and “spritzer” API are free. Also, SM sites have often changed access to their data through APIs for instance by introducing quotas. Some SM sites such as LinkedIn have almost completely shut down access through APIs.

A data mining task usually begins with understanding the domain. Opinions are expressed differently depending on if they are about political events, products or holiday destinations. So in the data collection phase not only the access methods have to be evaluated, but also the search terms have to be defined.

B. Data pre-processing

Raw data is seldom in a form that is useful for data mining. SM data is noisy, full of irrelevant information for analysis and contains a lot of spam. The data has thus to be cleaned, and *relevance filtered* first. Data cleaning is a time-consuming and labor-intensive procedure, but one that is absolutely necessary for successful data mining [9]. For opinion mining, only the phrases expressing the sentiment have to be extracted. Opinion mining is highly domain specific, and the first task is to define the sentiment words to look for. For instance an opinion can be expressed using sentiment word such as “great”, “excellent”, “awful”, using verbs such as “like”, “love”, for instance “the new iPhone is great” or “I like this car”. Sentiments can also be expressed using idioms such as “this car cost me an arm and a leg” or words that don’t hold a sentiment, for instance “this beer is flat”. Other common tasks in opinion mining are stop words removal, finding word stems using *stemming algorithms* and grouping the different inflected forms of a word so it can be analyzed as a single item using *lemmatization algorithms*. Once the sentiment words or phrases have been defined for a specific domain, the SM posts can be analyzed for their *sentiment polarity*. A list of sentiment words is called a *sentiment lexicon*, and these approaches are called sentiment lexicon based opinion mining.

ML algorithms usually don’t process text as input, they need a *feature vector*. Texts have to be represented in the vector space based on *Vector Space Modeling* (VSM). Feature vectors can be word frequencies of sentiment words, *part of speech* (POS) or sentiment polarity shifters, or *word weights*. *Term Frequency* and *Inverse Document Frequency* (TF-IDF) is one of the best known term weighting methods [23]. It is defined as:

$$w_{t,d} = tf_{t,d} \times \log\left(\frac{N}{df_t}\right) \quad (1)$$

where $tf_{t,d}$ is the number of occurrences of term t in the document d , N is the number of document in the collection and df_t is the number of documents, in which term t appears [23]. The posts are then classified according to their sentiment polarity based on their similarity using a distance measure such as the Euclidian distance, the Manhattan distance or the Chebyshev distance.

Another approach to creating inputs for ML algorithms is creating *bag-of-words*. A bag-of-words is a list of all the words in a text disregarding grammar or word order. They are often used when mining news articles for opinions, but can be applied to SM data too. Bag-of-words based approaches model news articles by vector space model which translates each news piece into a vector of word statistical measurements, such as the number of occurrences, etc. [22]. Bag-of-words are suitable as inputs for ML algorithms. They have the advantage that some of the data cleaning steps such as stemming or lemmatization can be omitted, however, they tend to perform less well when a lot of slang terms or special characters such as emoticons are used in posts.

C. Data mining

Data is mined to understand the underlying structure of the data and to make predictions based on historic data. It is the process of finding useful, actionable patterns in data and transform the raw data into knowledge. Opinion mining of SM posts is a *text classification* problem where posts are classified according to their sentiment polarity. SM posts can also be categorized using *clustering techniques* [24]. ML techniques are a suitable way for classification as well as clustering.

There are many ML learning techniques. They fall into three categories, *supervised*, *unsupervised* and *semi-supervised* ML models. Supervised ML techniques are used when the class label is known. For instance, when classifying Tweets into positive and negative Tweets, the class labels are “positive” and “negative”. Supervised techniques are used for classification and regression, unsupervised techniques are used for clustering when the class label is not known. Semi-supervised methods are used when there is a small amount of labeled data and large amounts of unlabeled data. For instance in genome sequencing there is usually a small sample size n and a large number of markers p , “large p small n problem”. Semi-supervised techniques can alleviate this problem [20]. The model is first trained using the small sample set, then it is applied to the large, unlabeled data set.

Ultimately we want to find a decision function f , which classifies SM posts according to their sentiment polarity. In the case of binary sentiment classification, we group posts into positive, P , and negative, N , reviews. If we denote the set of all posts by T , we search for a function $f:T \rightarrow \{P,N\}$. We use a random set of pre-classified training posts $\{(t_1,c_1), (t_2,c_2), \dots, (t_n,c_n)\}$, where $t_i \in T$ and $c_i \in \{P,N\}$ to train the learning scheme.

Experience shows that no single machine learning scheme is appropriate to all data mining problem [9]. Usually, several ML schemes are trained, and the one that has the best classification accuracy will be chosen. ML techniques include *naïve Bayes classifier*, *decision tree induction*, *Support Vector Machines* (SVM), *artificial Neural Networks* (aNN) and *k-Nearest Neighbor* (k-NN), but there are many more. They are well studied and have been applied in virtually any data mining domain. ML techniques such as aNN can handle very complex problems and give good approximations. However, they also tend to become complex themselves making it difficult to optimize. SVM are similar to aNN but are much simpler.

1) Support Vector Machines

Support Vector Machines (SVM) are based on *statistical learning theory*. SVM create a feature space or vector space defined by a *similarity matrix* (kernel) and create a hyperplane, an *affine decision surface*, to separate the training set. Support vector machines select a small number of critical boundary instances called support vectors from each class and build a linear discriminant function that separates them as widely as possible [9]. They maximize the distance from the closest training samples and transcend the limitations of linear

separations by including nonlinear terms and thus creating higher order decision boundaries. The techniques are related to the perceptron, which separates the training data set using a linear function. Perceptrons can be organized in interconnected layers creating a multilayer perceptron, an artificial neural network, to create a nonlinear decision boundary. Multilayer perceptrons allow to get approximations for very complex problems, however, they are complex in itself. SVM are a much simpler alternative and have become very popular in recent research.

If the training data is linearly separable, then a pair (w, b) exists such that:

$$w^T x_i + b \geq 1, \text{ for all } x_i \in P$$

$$w^T x_i + b \leq -1, \text{ for all } x_i \in N$$

with the decision rule given by:

$$\int_{w,b} (x) = \text{sgn}(w^T x + b) \quad (2)$$

where w is termed the weight vector and b the bias (or $-b$ is termed the threshold) [20].

SVM have been used primarily for classification, but they can also be used for regression.

2) Ensemble learning

Combining the output of several different models can make decisions more reliable. This process is called *ensemble learning*. Prominent methods include *bagging*, *boosting* and *stacking*. By combining several weak learning schemes, it is often possible to create a strong one. Ensemble learners have performed astonishingly well, but researchers have been struggling to explain why. For example, whereas human committees rarely benefit from noisy distractions, shaking up bagging by adding random variants of classifiers can improve performance [9]. Ensemble learning can comprise hundreds of models which makes it difficult to understand which factors improve the performance.

Probably the best performing ensemble learning scheme is *boosting*. Boosting combines models that complement each other. The models are of similar type, for instance, decision trees. Boosting iteratively builds models based on the performance of the last model such that the new model is trained on instances that were incorrectly classified by the last trained model. This only works well if each model correctly classifies a significant amount of data. Also boosting doesn't tread models equally but contrary to bagging weights a model's contribution by its confidence.

A boosting method designed specifically for classification is *AdaBoost*. AdaBoost calculates the weight of a model based on the models overall error e . The error rate is just the proportion of errors made over a whole set of instances, and it measures the overall performance of the classifier [9]. The weight w is then calculated as:

$$w = -\log \frac{e}{1-e} \quad (3)$$

Ensemble learners have many properties that make them

very suitable for SM data analysis. For instance models that identify spam with high accuracy such as the naïve Bayes classifier or perceptron [27] can be combined with models that are performing well in relevance filtering or classification thus creating a stronger learner than a single trained model.

Ensemble learners adopt a divide and conquer strategy in that they combine different learners with different accuracies in order to obtain a composite model that leverages the weakness of each single model. For example, *Instance Selection (IS)* is often used to handle noise [25]. Combining such a model with a model that is suitable for a specific classification problem can improve classification accuracy and also reduce the effort that goes into data pre-processing. SM data can thus be processed by different models, models that eliminate spam, models for relevance filtering and finally models for the actual classification.

Ensemble learners can handle very complex data mining problems, but they can become very complex themselves which runs counter to *Occam's razor*, which advocates simplicity. Loss of interpretability is a drawback when applying ensemble learning, but there are ways to derive intelligible structured descriptions based on what these methods learn [9]. Ideally instead of having an ensemble of learners, which makes it very difficult to interpret what kind of information has been extracted from what data, a single model would be preferred. If the ensemble learner is composed of decision trees, it is possible to combine them into a single structure, but it might still be difficult to interpret. An alternative are *LogitBoost trees*, which induce trees using *linear-logistic regression models* at the leaves. LogitBoost is an extension to the AdaBoost algorithm. It replaces the exponential loss of Adaboost algorithm to *conditional Bernoulli likelihood loss* [28]. If the LogitBoost algorithm is run until convergence, the result is a *maximum-likelihood, multiple-logistic regression model*. Running till convergence occurs is often not feasible due to performance issues when run against future, unseen data. However, it usually not necessary to wait until convergence to obtain good results. AdaBoost and LogitBoost are a very efficient classification method on *balanced data sets*. In real-world data, it is quite common to have *unbalanced classification data* and extensions to LogitBoost have been proposed [28],[29] to overcome this problem.

IV. CHALLENGES

Opinion mining remains a challenging area of research. Next to the regular challenges such as *sentence boundary disambiguation*, *word disambiguation*, and *sarcasm detection*, SM sites have certain properties which pose additional problems.

Spam has become a major issue on the Internet. Fake opinions are very difficult to detect, and opinion spammers often have fake identities (sock puppet, catfish).

Slang and jargon used in SM posts pose a major challenge for opinion mining. It is often specific to certain types of sites

such as dating sites, political discourse forums or product review sites. Also, many SM sites have specific characteristics such as the dollar sign denoting a company, e. g. “\$AAPL” for Apple Inc. or the hash tag “#” denoting the subject in Tweets. Abbreviations such as LOL (Lough out loud), IMHO (In my humble opinion) or AFAIK (As far as I know) are also widely in use, especially on microblogging sites where the number of characters per post is limited.

Noisy texts pose additional challenges since many ML algorithms such as naïve Bayes don’t handle it very well. Also, SM posts tend to be grammatically less correct and have many spelling errors which makes for instance sentiment lexicon based opinion mining or POS tagging less accurate. Often spelling errors are intended, for example for emphasis, e. g. “Gooooooooood camera”.

Most learning algorithms try to learn from noisy data by modeling the maximum likelihood output or least squared error, assuming that noise effects average out [26]. However, this method only works well for *symmetrical noise distributions*. Sources of noise in SM are typically asymmetrical, and many classification schemes such as naïve Bayes do not work well in these conditions.

SM site users decide themselves if they want to post an opinion on a certain subject, and the *self-selection bias* applies.

Identifying background topics that have been discussed for a long time and that are irrelevant to the public’s opinion is another issue that has to be addressed. Text clustering and summarization techniques are not appropriate for this task since they will discover all topics in a text collection [10].

Lastly, there are challenges inherent in ML techniques. Some models such as decision trees or aNNs tend to be overfitted. *Overfitting* occurs when the model captures noise instead of the actual opinion phrases. It happens when a model becomes too complex, and Occam’s razor applies.

V. CONCLUSIONS

Ensemble learners have worked surprisingly well when analyzing SM data. They are very robust also when data is noisy. However applying them requires a lot of experience and more research in certain areas is highly desirable. Making ensemble learners simpler by analyzing which features contribute to what extent to the result is one of the goals of our research. Ultimately we would like to have a learner that consists of only one model or at least only a few models with a clear separation of which model extracts what information. Simplifying models without losing predictive performance is an area where we would like to see more research effort.

Data pre-processing is an important step, and there seems to be much less research in data cleaning and feature selection than in the actual data analysis tasks. Spam or fake opinion detection remains difficult and more studies in this area could improve classification accuracy a lot. Feature selection is at least as important as selecting the most suitable learning scheme, and more research could lead to improved data

mining results.

Correlation doesn’t mean causation. If there is a correlation for instance between the number of positive Tweets and the sales volume of a product it doesn’t mean there is also a causal link. It is generally difficult to find the exact causes of sentiment variations since they may involve complicated internal and external factors [10]. A more holistic research approach could analyze the factors that influence positive reviews and product sales and lead to a clearer understanding of the causation.

Most studies treat every post equally. But some posts might be more influential because more people read them, or the poster has a higher authority. There has been some research on finding influential people in SM or in analyzing the online authority of users. Analyzing the impact of for instance opinion Tweets would improve opinion mining since some Tweets might be more influential because they have more followers or are more authoritative. SM posts could then be graded by their influence that would improve the predictive power of SM analysis.

REFERENCES

- [1] P. Włodarczak, J. Soar, and M. Ally, “Big Personal Data”, Social Science Research Network, 2014.
- [2] S. McGlaun, “Facebook data grows by over 500 TB daily”, SlashGear, 2012.
- [3] Twitter (2015, January), About, Available: <https://about.twitter.com/company>.
- [4] J. Bollen, H. Mao, and X.-J. Zeng, Twitter mood predicts the stock market, *Journal of Computational Science*, vol. 2, pp. 8, 2010.
- [5] S. Asur, and B. A. Huberman, Predicting the Future with Social Media, presented at the IEEE Int. Conf. Web Intelligence, 2010, pp. 492-499.
- [6] H. Achrekar, A. Gandhe, R. Lazarus, Y. Ssu-Hsin, and L. Benyuan, Predicting Flu Trends using Twitter data, presented at the IEEE Computer Communications Workshops (INFOCOM WKSHPs), 2011, pp. 702-707.
- [7] G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, V. Štrifský, and A. Holzinger, “Computational approaches for mining user’s opinions on the Web 2.0,” *Information Processing & Management*, vol. 50, no. 6, pp. 899-908, 11//, 2014.
- [8] N. Oliveira, P. Cortez, and N. Areal, “Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter,” in *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, Madrid, Spain, 2013, pp. 1-8.
- [9] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining*, 3 ed., Burlington, MA, USA: Elsevier, 2011.
- [10] T. Shulong, L. Yang, S. Huan, G. Ziyu, Y. Xifeng, B. Jiajun, C. Chun, and H. Xiaofei, “Interpreting the Public Sentiment Variations on Twitter,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 5, pp. 1158-1170, 2014.
- [11] P. Włodarczak, J. Soar, and M. Ally, “What the future holds for Social Media data analysis,” *World Academy of Science, Engineering and Technology*, vol. 9, no. 1, pp. 545, 2015.
- [12] M. Arias, A. Arratia, and R. Xuriguera, “Forecasting with twitter data,” *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 1, pp. 1-24, 2014.
- [13] B. Bechtolsheim, (2014, July) Google Cloud Platform is 11 for 12 in World Cup predictions. Google Cloud Platform. [Online]. Available: <http://googlecloudplatform.blogspot.ch/2014/07/google-cloud-platform-is-11-for-12-in-World-Cup-predictions.html>.
- [14] V. Shet, (2014, July) Microsoft’s Cortana predicts that Germany will win the FIFA World Cup 2014, sportskeeda. [Online]. Available: <http://www.sportskeeda.com/football/microsofts-cortana-predicts-germany-will-win-fifa-world-cup-2014>.

- [15] J. Tang, Y. Chang, and H. Liu, "Mining social media with social theories: a survey," SIGKDD Explor. Newsl., vol. 15, no. 2, pp. 20-29, 2014.
- [16] S. Stieglitz, and L. Dang-Xuan, "Social media and political communication: a social media analytics framework," Social Network Analysis and Mining, vol. 3, no. 4, pp. 1277-1291, 2013/12/01, 2013.
- [17] V. Hangya, and R. Farkas. Target-oriented opinion mining from tweets. in Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on. 2013.
- [18] D. Zlacký, J. Stas, J. Juhar, and A. Cizmar, "Text Categorization with Latent Dirichlet Allocation," Journal of electrical and electronics engineering, vol. 7, pp. 161-164, 05/01, 2014.
- [19] A. Kyriakopoulou, and T. Kalamboukis. Text classification using clustering. in Proceedings of The 17th European Conference on Machine Learning and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), Burlin, Germany. 2006.
- [20] Yip, K., C. Cheng, and M. Gerstein, Machine learning and genome annotation: a match meant to be? Genome Biology, 2013. 14(5): p. 205.
- [21] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2 ed., Heidelberg: Springer, 2011.
- [22] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," Knowledge-Based Systems, no. 0, 2014.
- [23] V. Hangya, and R. Farkas, "Target-oriented opinion mining from tweets." pp. 251 -254.
- [24] B. Liu, Sentiment Analysis and Opinion Mining: Morgan & Claypool, 2012.
- [25] S. B. Kotsiantis, "Supervised Machine Learning," Informatica, vol. 31, pp. 19, 2007.
- [26] M. D. Schmidt, and H. Lipson, "Learning noise," in Proceedings of the 9th annual conference on Genetic and evolutionary computation, London, England, 2007, pp. 1680-1685.
- [27] K. Tretyakov, "Machine Learning Techniques in Spam Filtering," Data Mining Problem-oriented Seminar, U. o. T. Institute of Computer Science, ed., 2004, p. 19.
- [28] S. Jie, L. Xiaoling, L. Miao, and W. Xizhi, "A new LogitBoost algorithm for multiclass unbalanced data classification." pp. 974-977.
- [29] J. Song, X. Lu, and X. Wu, "An Improved AdaBoost Algorithm for Unbalanced Classification Data ". pp. 109 - 113.

Design of methodology for connecting Enterprise Architect with development solutions and necessary application framework

J. Sedivy, R. Borkovec, P. Coufal

Abstract— The basic idea of this article is to optimize the development work in order to maintain the quality and purity of workflows and their developer, professional level. A reliable implementation base is called. Framework. Framework (Application Framework) is a software architecture That Serves as support for programming and the development and organization of other software projects. This May include support programs, libraries, API, support for design patterns and best practices for development.

Keywords— enterprise architect, framework, implementation base framework, UML language.

I. INTRODUCTION

WHEN we implementing software contracts we have to respect these fundamental points [1]:

- a) professional approach to client issues
- b) reliable execution platform that will be common to all projects implemented
- c) the creation of high quality, accurate documentation, both technical and user
- d) ensure the possibility of product updates and changes with minimal need for recompiling and builds
- e) the open interfaces for multi-language solutions

II. PROFESSIONAL APPROACH TO THE PROBLEM OF CLIENT

If the editing software procurement incomplete or inaccurate, even the best application can not reach a satisfactory outcome. Prerequisites for that award went well and gave so much developer indispensable basis for their work, could be summarized in the following points:

- a) knowledge of solved problems

This article was created under the project called Specific research done at UHK Hradec Králové in 2015.

J. Sedivy, University of Hradec Kralove, Faculty of Science, Department of Informatics, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic (phone: +420 493331171; e-mail: josef.sedivy@uhk.cz).

R. Borkovec, University of Hradec Kralove, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic (phone: +420 493331171; e-mail: roman.borkovec@uhk.cz).

P. Coufal, University of Hradec Kralove, Faculty of Science, Department of Informatics, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic (phone: +420 493331171; e-mail: petr.coufal@uhk.cz).

b) the communication skills of the analyst, establish a good communication channel with the client and use those resources that will actually use the resulting application

c) respecting the connection to the outside world, whether it is necessary to import and export data, or links to other applications [2]

d) to estimate the direction of further development or degree of autonomy applications

Of those paragraphs that would be at this stage of development analyst paid tool that allows him to communicate with the client at the earliest opportunity to view his ideas in the resulting environment [3]. On the client can not assume a rigorous understanding of UML schemas, even more can be expected that this ignorance and admit because there is a misunderstanding. Which is where the award is a fatal error. How to solve this conflict, suggesting another subchapter.

III. RELIABLE IMPLEMENTATION BASE FRAMEWORK

The aim of the framework is the assumption of the typical problems of the region, thereby facilitating the development so that designers and developers can focus on their task. For example, a team that uses Apache Struts to develop websites for the bank can focus on how they will carry out banking transactions and not to ensure flawless navigation between pages.

There are objections to using the framework code will slow or otherwise ineffective and that the time is saved by using foreign code must be given to staging framework. However, when his repeated commitment or large project, there is a substantial time savings. When uninstalling framework will not be able to run some applications.

Framework consists of the so-called. Frozen spots and hot spots. Frozen spots define the overall architecture of the software structure, its basic components and the relationships between them. These parts do not change in any way the use of the framework. In contrast, hot spots are components that together with the programmer code creates a very specific functionality and therefore are almost always different [4].

In object-oriented environment framework consists of abstract and conventional classes. Frozen spots can then be represented by abstract classes and custom code (hot spots) was added to the implementation of abstract methods.

Examples:

- JUnit is a framework for testing the units for the programming language Java.
- Spring is an application framework for the Java platform open source.
- Zend Framework is a framework for web applications in PHP open source.
- Vaadin is a framework for web applications in Java open source.
- Nette Framework is a framework from Czech author for web applications in PHP open source.
- CakePHP Framework is a framework for web applications in PHP open source.
- Symfony is a framework for developing web applications in PHP open source.
- CodeIgniter is a framework for developing web applications in PHP open source.
- Apache Wicket is a framework for developing web applications in Java open source.
- Ruby on Rails is a framework for web applications in Ruby open source.
- jQuery is a lightweight JavaScript framework with open source code.
- The .NET Framework is a framework for language C #

In our particular case is not so much about ORM (Object Relational display O / RM or O / R display) is a programming technique in software engineering, which provides automatic conversion of data between relational databases and object-oriented programming language.), or similar solutions. It is a comprehensive system that meets all specified requirements. The basic idea of the whole framework is a generic creation majority of forms and direct modifications without having to build the application itself [5].

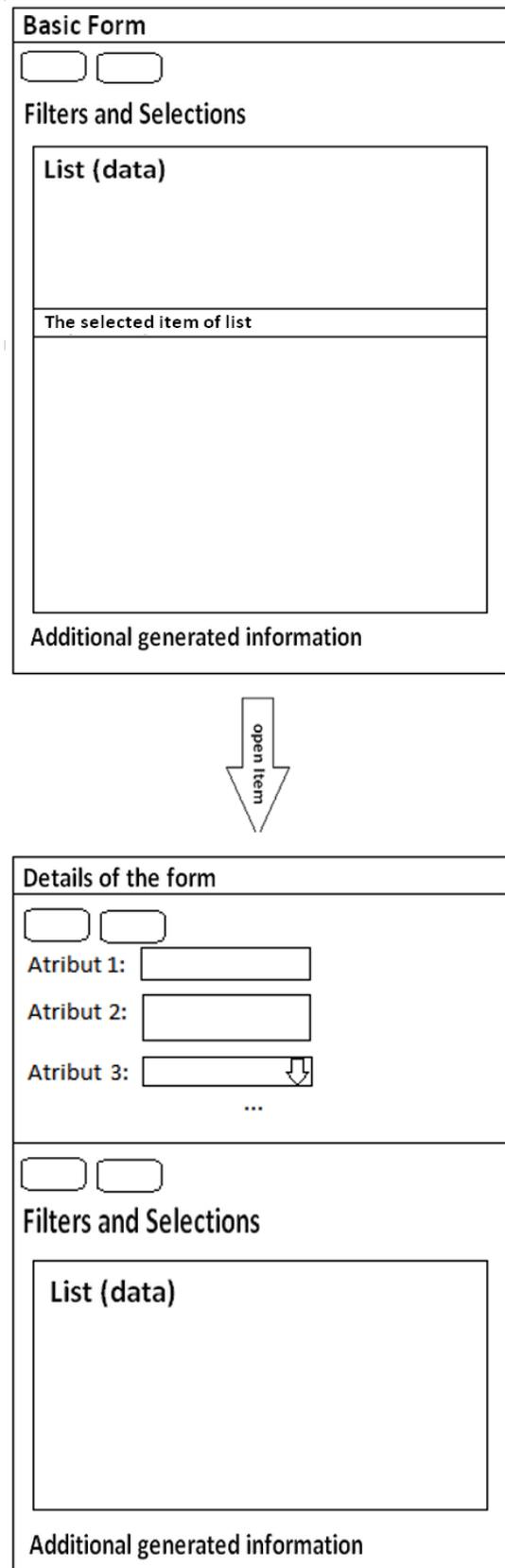


Fig. 1 Majority form

The dependent lists detail the course progresses further recursively, where possible approaches and all things related to the access and business logic of the application.

On the form, in the upper part of the visible buttons that allow you to extend the functionality of form, they can be called up outside, programmer-designed elements - whether it is a form of executive or call stored procedures on the database server or specific methods of application code - in Then use .NET reflection in its entirety. In the case of detail is a suitable place for the location type buttons Save, Undo, Copy form, etc.

By being available to the interface, it will receive only a single ergonomic presentation layer, but also an interface that is effective at the time of application design with the contracting authority [6]. Receives direct vision of the future appearance, layout elements (whether columns in the list, the filtering, sorting, but the arrangement detail forms, including selection of appropriate components - calendars, vintage elements, check elements etc.)

Definitions of these elements can be implemented at least two ways - either by direct description in the database where there is both data and business layer and also a description of the presentation. This method allows to realize the most fundamental changes, such as bringing not only development, but also for future use applications directly, without having to build applications. Or, in the case of a consistent three-layer application can be used as source elements class that reflection transformed into the final application solutions. In this case, it is necessary when a change in classes with a build count [7].

By drafting the basic design solutions in collaboration with the client, creating the necessary elements in the database. These can be used for transformation using the Add in EA generate analytical model so that it remains in direct connection with the application. All changes will be reflected on both sides, both from the model to the application, and vice versa.

IV. CREATION OF QUALITY DOCUMENTATION UNITS

Technical documentation is perfect solved on the level of development tools, both Visual Studio, as well as the actual Enterprise Architect [8]. But remains the issue of user documentation, which does have the normal function of the sporadic use - user rarely actually use this documentation, but it is absolutely necessary part of the assembly task.

Proposed Framework enables a solution that ensures due to the following functions:

- Reducing the overhead required for creating user documentation that is partly generated and partly it has the ability to directly generate user
- Differentiation between the different parts of the dossier by access roles. It means that there is no need to provide the resulting complete user manual, which is in charge of a strict application space. Documentation is generated only for those elements that for him the privileges accessible

- Use of framework options for printing and folding the proper coverage documentation

V. LINKS TO THE SURROUNDINGS, EXPORT - IMPORT DATA PRINTS

Component Solutions Framework allows you to gradually expand its capabilities, and this also applies to the issue [9]

The basic component that will provide these options is a list. After narrowing the data and their arrangement, the user will want this data to be exported outside the application. The first step in the implementation is likely to be linked to Microsoft Office tools, which can be connected to their templates to use very widely. In some applications, this solution is quite sufficient and does not need to dial another extension.

A good step is to ensure communication with the Open Document protocol and XML - in a generic and component solutions is an advantage not only time-space, but also full applicability throughout the application in those places where the import or export offer.

Of course, in specific cases, intervention can not be avoided in the application code, but for standard situations, all of these interventions can be defined directly in the database, or builds again without updating the application itself.

VI. OPEN MULTI-LANGUAGE INTERFACE APPLICATION UPDATE

Today, the requirements for localized increase in proportion with the growth rate of transnational projects and also in providing information and services outside one specific destination [10].

Question localization is not just a question of translation output strings. There can be addressed relatively simple function, which also ensures the fulfillment of the phrase in the dictionary itself with the application. It is also addressing local codebooks and possible differences in some procedures, eg. Calculation or legal.

An important issue is to provide a simple interface for translators and the possibility of pairing their work. The translator can not be bother complex operations, but it is necessary to ensure the synchronization of their work and the maximum overall efficiency and flawless.

VII. CONCLUSION

Options and concrete solutions are very wide and after each team has to learn basic functionalities opportunity to constantly replenish their framework and improved. The basis should always be the effectiveness of development, optimizing data access and maximum user friendliness. In an environment. NET is worth solve database level using DB Providers - then there is no need to develop two frameworks just because the company develops applications eg for Firebird and MSSQL or Oracle. Basic core framework is thus independent of the database, and it enables you to provide a wider client access,

which in turn are inputs into other financial firms.

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

ACKNOWLEDGMENT

This article was created under the project called Specific research done at UHK, faculty of Science, Hradec Králové in 2015.

REFERENCES

- [1] P. Voborník, Migration of the Perfect Cipher to the Current Computing Environment. *WSEAS transactions on information science and applications*. 2014, 2014(11), s. 196-203. ISSN 1790-0832.
- [2] M. Musílek, Computer Aided Teaching Topic "The Rainbow Formation" in Subject Computer Simulation in Physics at High Schools. In: *DIVAI 2014: 10th International scientific conference on distance learning in applied informatics*. Praha: Wolters Kluwer ČR, 2014, s. 431-440. ISBN 978-80-7478-497-2.
- [3] P. Voborník, Mini-Language for Effective Definition of the Color Gradients. In: *Materials, transportation and environmental engineering II (CMTEE 2014)*. Zurich: Trans tech publications, 2014, s. 1882-1885. ISBN 978-3-03835-248-8.
- [4] R. Němec, M. Hubálovská, Š. Hubálovský, User Interface of System SMPSL. In: *Communications and information technology (CIT 2014)*. Salem: North atlantic university union, 2014, s. 324-329. ISBN 978-960-474-361-2.
- [5] Š. Hubálovský, M. Musílek, Algorithm for Automatic Deciphering of Mono-Alphabetical Substituted Cipher Realized in MS Excel Spreadsheet. In: *Applied science, materials science and information technologies in industry*. Zurich: Trans tech publications, 2014, s. 624-627. ISBN 978-3-03835-012-5.
- [6] P. Voborník, Modification of the perfect cipher for practical use. In: *Manufacturing, engineering, quality and production systems (MEQAPS 2014)*. Athens: World scientific and engineering academy and society, 2014, s. 64-68. ISBN 978-960-474-387-2.
- [7] Š. Hubálovský, Modeling and Computer Simulation of Static, Dynamic and Feedback Systems as Tool of Development of Logical Thinking. *International journal of mathematics and computers in simulation*. 2014, 8(2014), s. 276-285. ISSN 1998-0159.
- [8] E. Milková, A. Hůlková, Algorithmic and Logical Thinking Development: base of programming skills. *WSEAS transactions on computers*. 2013, 12(2), s. 41-51. ISSN 1109-2750.
- [9] Š. Hubálovský, Modeling and Simulation of Real Process – Passing through the Labyrinth as a Method of Development of Algorithm Thinking and Programming Skills. *International journal of mathematics and computers in simulation*. 2013, 7(2), s. 125-133. ISSN 1998-0159.
- [10] E. Milková, Development of Algorithmic Thinking and Imagination: base of programming skills. In: *CSCC 2012: proceedings of the 16th WSEAS international conference on communication*. Athens: World scientific and engineering academy and society, 2012, s. 347-352. ISBN 978-1-61804-109-8.

Ing. Mgr. J. Sedivy, Ph.D., was born in 1963 in Czech Republic. Doctor degree in Theory of technical education in 2006 on University of Hradec Kralove, Faculty of Education, Czech Republic. University of Hradec Kralove, Faculty of Education, Department of Technical Subjects, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic (phone: +420 493331171; e-mail: josef. sedivy@uhk.cz). His scientific activities are computer graphics and communications in education and informatics.

Ing. R. Borkovec, was born in 1965 in Czech Republic. System Engineer on University of Hradec Kralove, Czech Republic. University of Hradec Kralove, Faculty of Education, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic (phone: +420 493331171; e-mail: josef. sedivy@uhk.cz). His scientific activities are programming, education and informatics.

Bc. P. Coufal, was born in 1992 in Czech Republic. Bachelor degree in Informatics Education on University of Hradec Kralove, Faculty of Education, Czech Republic. University of Hradec Kralove, Faculty of Science, Department of Informatics, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic (phone: +420 493331171; e-mail: petr.coufal@uhk.cz). His scientific activities are programming, robots, informatics education.

New Challenges in Smart Campus Applications

Attila Adamkó, Tamás Kádek, Lajos Kollár, Márk Kósa, János Pánovics

Faculty of Informatics

University of Debrecen

Kassai út 26, H-4028 Debrecen, Hungary

Email: {adamko.attila,kadek.tamas,kollar.lajos,kosa.mark,panovics.janos}@inf.unideb.hu

Abstract—Nowadays very common keywords are **Big Data**, **IoT** (Internet of Things), **crowdsourcing** and **ubiquitous computing**. All of them gained greater emphasis and our University Campus is a great place where all of these areas could be investigated. Wide ranges of data could be collected from the built-in sensors of the building and naturally, from the users smartphones or tablets resulting a huge amount of data.

On one hand, our paper includes a framework which could provide value-added services for various people living or working on the Campus. On the other hand, the Campus is a perfect place where new algorithms could be developed and tested through these services. Furthermore, the community could be involved not just as subscribers for the services but also as providers of the data, and in an optimal case, the crowd could prepare and provide new information sources.

Index Terms—campus, smart, adaptive, intelligent systems, crowdsourcing

I. INTRODUCTION

OUR primary goal was to create an architectural framework which allows various members of the community to create and use services based on the data that is collected in a university environment. These data include information on course enrollments, timetable, exam dates, office hours, and various deadlines along with community provided data. These collected data can then be subject to analysis, based on which they can either be fed back into the services, or we can provide recommendations or offer new services for our users.

Future Internet research—which was appeared at least seven years ago—aims at bridging the gap between both academic and industrial community’s visionary research and large-scale experimentation [11], [2].

One part of it is the Internet of Things (IoT) phenomenon which highlights the opportunities lying in the sensors connected through wireless connections. We have successfully applied it one of our scenarios where the location of the users is crucial. Moreover, these sensor data serve as an endless source of environmental data which could drive a real-time and/or a transactional analytical module. In our vision this could be used to create trajectories and help users to find nearby colleagues or friends based on historical presence data and realtime sensor information.

The next piece of the trends is expressed by the term of ubiquitous computing (ubicom) which states that computing could appear anywhere and everywhere. Borders are blurred between computing devices including desktop computers, notebooks, tablets and smartphones. Our framework currently

includes a mobile and a desktop version which makes available the seamless and smooth usage of the system. Naturally, it is not mean the interface is the only thing which need to be available on the different devices. User profiles are created to support context-aware and customized environments for the ongoing research.

The third pillar of our framework is the crowd. Crowdsourcing could be applied in a University Campus as there are lots of people (students and staff) with different interests and different requirements for the services. However, they are not only consumers of the information produced by the system, they are producers—and content generators—as well. Lots of data are generated while various applications are used by them.

II. SMART COMMUNITIES

In this paper, when we explain the concept of the smart community we concentrate on the expectations connected with the software operating in a community. In our approach, the smart community is such a community that is served by smart applications. Based on the definition of the “Apps for Smart Cities Manifesto” for Smart Cities [7], the requirements of the smart community applications are the following:

- sensible—the environment is sensed by sensors;
- connectable—networking devices bring the sensing information to the web;
- accessible—the information is published on the web, and accessible to the users;
- ubiquitous—the users can get access to the information through the web, but more importantly in mobile any time and any place;
- sociable—a user can publish the information through his social network;
- sharable—not just the data, but the object itself must be accessible and addressable;
- visible/augmented—make the hidden information seen by retrofitting the physical environment.

An application itself, which satisfies partly or fully these requirements, could not be called smart at all. The basic requirement of a smart service to have information about the community. First of all, we have to collect and publish the available information. At the current level of technical development, the collected information is very huge and heterogeneous. A smartphone with average performance is also capable of GPS-based localization, light detecting, making photos, detecting of mobile and wireless network devices, etc.

Of course, the sensors are not always in the devices of the end users, think about of, e.g., a handle for free parking spaces for vehicles, a digital temperature sensor on buildings or an electronic passing gate system using RFID cards. These are all such kind of information which could be used to drive the creation of new services.

In many cases, the problem is that we have too much information. These information, temporarily, must be stored on the device that collects the data. The goal is, of course, the publishing of these data as soon as possible. The stage of the publishing is the web. In this place, we do not explain that it is practical to aggregate and to filter the collected data, moreover sometimes this procedure is mandatory due to the supporting of the anonymity [3]. Here let it be enough that the collected data must be maintained in a centralized, or, moreover, in a unified way, because of these data make the base of the smart services up.

The smart community is not static, it is continuously changing, and it is persistently on the move. The aspect and the amount of the collected information is changing from time to time. So, the evolution of the smart applications is a neverending process. New applications could be created, and their pure existence—the information about how we could use them—also could improve the amount of the sensible things. With this, we could collect new data about the working of the community. From the new data, we could know something about the behavior of the community, again, and these data could be reused in the life of the community.

III. THE EXTENSIBLE ARCHITECTURE

From the Smart Campus perspective, one of the main challenges is to collect content from various sources where the majority of them might be created at a later time. That is why we need a highly extensible system which is designed for change.

Such an extensible architecture has been designed and published by the authors in [1]. It allows the extension of the system with new elements on both the data producer and consumer side. This is where the crowd could help us by adding new sources and new services with the development of their own information parsers.

According to [1], a Smart Campus environment has lots of various (and most importantly, heterogeneous) data sources including the following:

- an Education Administration System called Neptun that contains information on course enrollments, timetable information of courses, exam dates and times, etc.,
- faculty members offering office hours, consultations, etc.,
- Education Offices of the various faculties offering office hours,
- Student Governments organizing events for students,
- the menus of the canteens located at the Campus,
- geolocation (e.g., GPS), WiFi or some other sensor data collected by smartphones or similar devices,
- data gathered by environmental and building sensors (temperature, humidity, air pressure, air pollution, etc.),
- a Library Information System that is able to tell whether a given book is available or not,

- social media sites (like Facebook or Google+) containing information on friends and ranges of interests of a person,
- professional sites (like LinkedIn) holding data on work experience and professional achievements (however, this is not necessarily the most important data source from Smart Campus perspective),
- bibliographic databases (like Google Scholar, DBLP or Scopus) that provide information of published journal articles or conference papers of researchers,
- event hosts of actually any events (like public lectures, concerts, exhibitions or whatever users might be interested in), and, which is essential,
- the crowd itself with the added value of the capability of generating content that is interesting for a set of people (or, to be more precise, consumers).

These are only examples of data sources not an exhaustive list. These demonstrate that what kind of diversity in data sources should a complex application face. Applications that provide value-added services typically require integration of some data coming from more of these data sources. That was the reason of developing an architecture providing the ability of accessing information from existing sources along with making the addition of new sources possible and (relatively) easy.

Some of the data can be collected in an automated way (e.g., sensor data), some others might require manual interaction (like canteens' menus or office hours of instructors); some of the data sources offer Application Programming Interfaces (APIs) to provide access to data (e.g., social media sites) while others do not have APIs therefore web crawlers are needed to gather and parse the data; some of the data sources provide built-in notification mechanisms (e.g., an event feed of a social network site) while others do not (for example, adding new office hours or changing the daily menu).

We have chosen the Extensible Messaging and Presence Protocol (XMPP) as the underlying communication protocol [9], due to its extensibility and publish/subscribe model. Further considerations on the design of the architecture are described in [1].

The power of our architecture as it does not limit the possible data sources and also allows the collection of some specific information. For example, if a couple of students prefer to have a lunch at the small restaurant near the Campus they can develop a connector that parses the restaurants web page to provide information on the daily menu. The same case can be also true for news feeds. When these sources are became available the potential (interested) users could be notified about it based on the preferences and the meta information provided for the feeds.

The Smart Campus Central Intelligence (SCCI) component in our architecture provides an interface between the information sources including both the incoming events (XMPP server) and the information stored in the database and the Web services layer (Figure 1).

It provides a couple of unified data models related to some of the most notable domains in a life of a Campus: educational data, research data, social data (with friend of and classmate of relationships), etc. SCCI layer also allows to define and

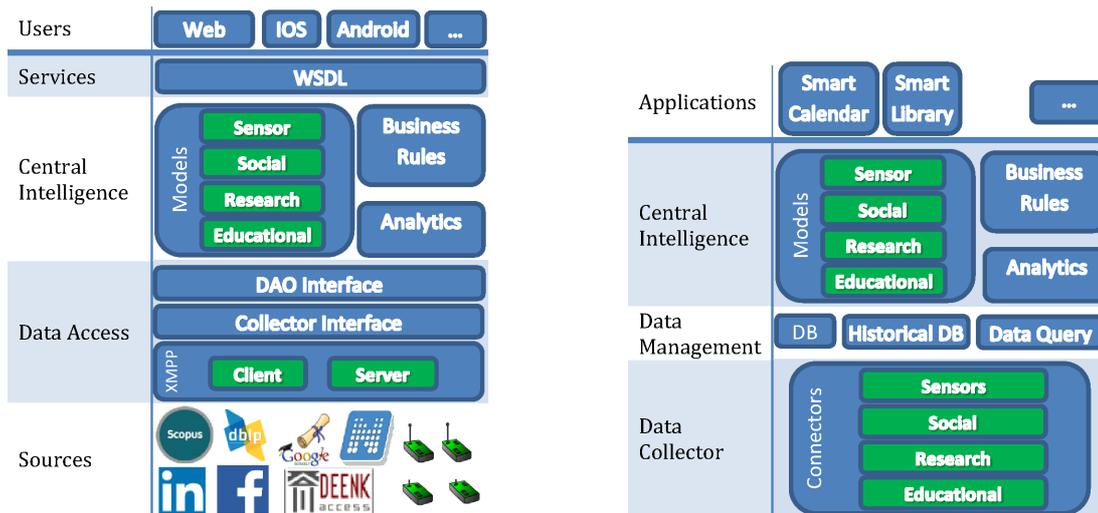


Fig. 1: Smart Campus Layers and Services

validate business rules and this is where the Analytics module resides. Now we have only a couple of simple analysis that does not exceed the scale of a medium-level database query but deeper analytical capabilities are planned to be added as this is how we can provide better value-added services.

An example where our services are created and used by the Smart Campus is the Faculty's portal which is based on Liferay. Its architecture is composed of three layers: Enterprise layer, Service layer and Persistence layer. While the Enterprise layer is responsible to fulfill all the enterprise needs (content, workflow, document, user, etc. management) the Service layer is the core component. It contains the majority of the business logic that perform enterprise needs. Liferay follows a Model Driven Architecture and this is made possible by the implementation of Service Builder which is used to automate the creation of interfaces and classes for database persistence and a service layer. In Liferay Web services has two mostly used and significant protocols: JSON Web Services and SOAP.

We have successfully implemented services in the portal to serve the lecturers' opening hours. One connector is used inside the Educational part to periodically check the available data. If there are changes the SCCI is notified. Based on the actual data all the affected users are notified about the change. Currently we are working on the extension of this portal service to alter it to a fully functional XMPP client. After that our portal service can directly notify the SCCI about the changes and the requested operations could be done without requiring the intermediate crawler.

An other ongoing development is the creation of a notification portlet that is integrated to the Faculty portal where users can see each other's online presence and start conversations without leaving the site. This is the advantage of the XMPP protocol and the usage of LDAP. It is straightforward because we can provide a platform where all the education-related tasks could be done therefore users can experience the added value of the services. Moreover, messages are not only sent between two users, but as an extension it is possible to have SCCI be a friend of all users. This gives the possibility of

notifying users within the portal system, without the need of the execution of any third-party clients. It is similar to what we have as push notifications on the mobile platform. The portal changes its information provider role to an online collaboration environment.

A. The life cycle of intelligent services

From our point of view, the basis of the intelligent community is to make the services online available. The first step is to collect the useful information into one online accessible database. First, it means a simple data service. For example, it could be a simple collection of news, a schedule, or even a timetable. In the second step, we have to pay attention on the use of available information. Notice the most popular news or columns, and then we can reorganize the information such as order the most popular columns to the first page. It is still not an intelligent service, but if we have a lots of application working together, which can share the information in the mentioned way, the service goes to be smarter and smarter.

IV. BASIC ELEMENTS FOR ADAPTION

Adaptive systems have gained increased popularity in the last decade. The overall goal is to improve the usability of the system and provide better user experience by applying personalization based on the services discussed in the previous section.

In a more technical view, we need to study and understood the structure of the metadata (i.e. the semantics) at the content's side which could fuel the adaptation process based on the user profile. Semantic markup included in the whole process with machine-understandable representation of the content. The work presented in this paper based on the previously mentioned extensible architecture where the Data Management layer contains the databases that are populated with data gathered by connectors of the underlying layer (see Figure 1). As of today, due to the characteristics of the data we have both a relational and a graph database as a backend. Well-structured information (e.g., course and timetable information)

are stored in a normalized relational database. However, for semistructured or unstructured information the rigid structure of relational tables are not appropriate, so we have selected neo4j as a solution.

The result is a highly semi-structured system. It is not just collecting large volume and variation of data but also understanding the relationship between entities by mapping their connection into our system. With this method we established a data store which can be easily extended. When a new—and previously nonexistent—source is attached to the system the only task we need to perform is to add the new nodes and proper edges to the database.

In this context, following the Semantic Web initiative [10], ontology-based annotation helps us to provide adaptive processes, meaning that the incoming content enriched by semantic data. One could imagine all of these as tags attached to the data which could be a JSON message, an HTML fragment or a Web Service call. We need to transform this data into triples. The available prefixes are defined in the ontology into which these triples are going to be imported. Traditional SPARQL could be used to made queries, but we have found that graph databases also could be applied for reasoning.

Along with the semantic tags, the next important piece in our system are the groups. These groups modelling the subscribers which are ground for the original architectural idea that based on the publish/subscribe model. Groups are high level entities and some of them are automatically created, like the group for a user's personal calendar or for the news feeds, etc. As a member of a given group you will receive the events published by that source, and naturally one user can join as many he/she wants and one may create new ones too. The visibility of the groups also could be controlled, there could exist public and private groups where freely everybody or only the invited members could join.

V. SMART CAMPUS APPLICATIONS

A. Adaptive Event Recommendation as a Personal Calendar

The first application which appeared in the concept of the Smart Campus at University of Debrecen was a simple data serving application gathering into one location all the important events at the Faculty. The first principle—following the outlines of the Intelligent application's lifecycle—was only to provide a uniform Web Service interface to made accessible all of that data. It could open the way to all of the end-user services, like reminding for important deadlines or browsing the categorized events.

However, collecting the information from distinct information systems and providing them through a standardized way is a useful approach—but cannot be seen as an intelligent one. On the contrary, the usage scenarios of the published information could serve as a base for predicting their behavior. Take the example, users could mark events as important to create their personal calendar. The system could analyze these marks and highlight the most frequent and important events. While checking the marks the system could detect that some of the events are closing to the physical limit of the room where it is scheduled. In that case, searching for a greater room

is essential and required. Finally, the system could notify all the attendees from the change of the location. Naturally, the notification is based on the previously mentioned group-based solution.

In that calendar we have developed the possibility to browse, categorize and register for events. Hereafter, we made available to rate those events. The service is available on the following address: <http://smartcampus.hu>.

An Android application (see Figure 2) has also been implemented to support easier access of those services. This mobile application allows additional services to use. As we have stated earlier, the base of an intelligent service is not just the direct information posted by the users but may origin from sensors as well. We investigated the possibilities in our building and prepared an application for smartphones which could determine with a very good approximation the position of the user based on the WiFi network access points. When connecting this data with the user's personal calendar, we could show the nearby events in the first place on the list. The demonstration of this service is planned for a local conference held in this autumn where the application will be used to show the nearby sections program at first place.

Moreover, the event recommendation system is not based just time and place attributes. The categorization made it available to recommend events for the user based on its topics. Imagine the situation when you mark important events and the system recommends you upcoming possibilities based on the categories—and (ontology based) related categories—of the marked events.

B. Adaptive Meeting Planning

For the above mentioned calendar we have an ongoing development to extend it with a meeting planning module. That module will read all the participants calendar events and try to suggest time slots which could be appropriate for all the attendees.

The adaption is based on properties of the event which is related to the user as well. It could be mandatory, optional or rescheduleable—like a registration for an opening hour but the meeting may be scheduled to the same time because one can attend an other opening hour at a later time [6].

C. Managed Programming Contests

The basic goal of the ProgCont system was to support the organization part of programming contests. The development process started in 2011 with a web application and worker services. The web application stores the exercises in a problem catalogue and collects the submissions including not only the solution source code, but the necessary information about the competitors and contests. The worker services are used to evaluate the submissions, they compiles them (if it is possible), and validate the compiled algorithm by running several test cases.

Of course the software itself could not be called smart at all, even if most of the exercises were selected from international programming contest for the purpose of exercising our students. But the ProgCont system collects numerous additional

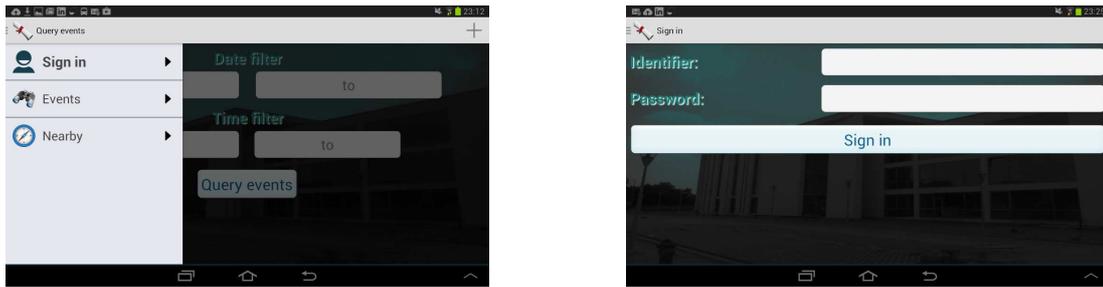


Fig. 2: Smart Campus Android application

data during its operation, which give us the ability to discover extra knowledge for the users [8].

One trivial example based on the success rate of submissions, more precisely the number of users who can solve the exercise, and the number of trials before the first successful solution. Everyone knows own success rate. The additional information in that case means that we made the cumulated results available.

Another useful information could be the quality measure of the accepted solutions. It means, that we measure and rank the accepted solutions using another criteria such as the length of source code, the required amount of memory, the execution time, the number of iterations, and so on.

There exists several another way to support our student with the ProgCont system. A catalog of thematic exercises are also available, giving the possibility to browse problems with the same algorithmic background. ProgCont can provide automatic suggestions if the necessary amount of information is available about the user, in other words, when there is enough evaluated submission to approximate the user skill.

One objective way to measure the worth of an exercise is to calculate the rate of accepted and unsuccessful submissions. But in the other hand, the users can also rate the exercises. Both ways can serve useful information for further decisions about how should the user continue the preparation process.

Nowadays the system also used in classroom test. The collected data gives the ability to do statistical investigation from the pedagogy point of view. Relative frequency histogram and relative frequency of scores are available according to the tested groups, and the connection between separate classroom tests can be discovered using a diagram representing correlation.

D. Assessment system for non-graded exercises

A Spring-based Web application with a JavaServer Faces (JSF) frontend has been developed in order to help course instructors in offering optional exercises or assignments (which do not count into the grade) for their students. In principle, these exercises can be of any kind (programming, database, math, etc.) but we primarily focused on programming exercises. The primary difference from ProgCont is that unlike ProgCont this system does not require any feedback from the instructor side. It was designed in order to decrease the load on instructors by asking the crowd (i.e., other students) to validate

submissions (that is why it is only used for optional exercises). Unfortunately, there are some courses at our institute where we have relatively few course instructors. Their number is enough for creating and validating assignments that count into the grade but short for doing the same for not graded exercises (e.g., they did not have enough time for preparing appropriate test cases in order to validate with ProgCont).

The Web application offers the following major functionalities. It is possible

- to define exercises,
- to submit solutions,
- to assess submissions,
- to comment and/or rate assessments.

Any registered user (i.e., even students) can define exercises. Those who solve an exercise can submit the solution which can be evaluated (assessed) by anyone else. Therefore practising students can receive some feedback, even if these answers cannot be trusted. To deal with false positive feedbacks, it is important that not only the solutions but the people giving feedbacks should also be rated. Later, those people's ratings who regularly give wrong answers will count less.

Similarly to ProgCont, the data gathered by this application can (and should) also be analysed using data mining techniques. Currently, assessments are created on a voluntary basis. However, good assessments will come from qualified people who understand several aspects of the submitted solution. It is very hard to find suitable reviewers whose knowledge level and experience is sufficient for providing valuable assessments. After analysing the gathered data we can classify possible reviewers based on how valuable their assessments are therefore the system will be able to offer a set of possible reviewers for each task.

VI. A SAMPLE CALENDAR SERVICE PROBLEM

Some of the problems that arise concerning Smart Campus applications might be solved using artificial intelligence methods. In this chapter, we present a sample problem related to the calendar service. In order to define the problem, we first introduce the Extended State-Space Model, then a state-space representation of the problem is given. Later, we summarize the Extended Breadth-First Algorithm (EBFS).

A. The Extended State-Space Model (ESSM)

The EBFS algorithm can be defined after introducing an extended state-space model [4], [5], which allows us to discover

the representation graph starting from several different states and possibly in more than one direction. Using state-space representation, solutions to problems are obtained by executing a series of well-defined steps. During the execution of each step, newer and newer states are created, which form the state space. States are distinguished from one another based on their relevant properties. Relevant properties are defined by the sets of their possible values, so a state can be represented as an element of the Cartesian product of these sets. Let us denote this Cartesian product by S . Possible steps are then operations on the elements of S . Let us denote the set of operations by F . The state space is often illustrated as a graph, in which nodes represent states, and edges represent operations. This way, searching for a solution to a problem can be done actually using a path-finding algorithm.

We keep the basic idea (i.e., the concepts of states and operations on states) also in the extended state-space model (ESSM). The goal of this generalization is to provide the ability to model as many systems not conforming to the classical interpretation as possible in a uniform manner.

A state-space representation over state space S is defined as a 5-tuple of the form

$$\langle K, \text{initial}, \text{goal}, F, B \rangle,$$

where

- K is a set of initially known (IK) states, such that $K \subseteq S$ and $K \neq \emptyset$,
- $\text{initial} \in \{\text{true}, \text{false}\}^S$ is a Boolean function that selects the initial states,
- $\text{goal} \in \{\text{true}, \text{false}\}^S$ is a Boolean function that selects the goal states,
- $F = \{f_1, f_2, \dots, f_n\}$ is a set of “forward” functions, $f_i \in (2^S)^S$,
- $B = \{b_1, b_2, \dots, b_m\}$ is a set of “backward” functions, $b_i \in (2^S)^S$.

The “forward” and “backward” functions represent the direct connections between states. For more details, see [4].

Some notes:

- The number of initial and goal states is not necessarily known initially, as we may not be able to or may not intend to generate the whole set S before or during the search.
- The $n + m = 0$ case is excluded because in that case, nothing would represent the relationship between the states.
- Although the elements of the sets F and B are formally similar functions, their semantics are quite different. The real set-valued functions in F are used to represent nondeterministic operators, while there may be real set-valued functions in set B even in case of deterministic operators.

Let us now introduce a couple of concepts:

- *Initial state*: a state s for which $s \in S$ and $\text{initial}(s) = \text{true}$.
- *Goal state*: a state s for which $s \in S$ and $\text{goal}(s) = \text{true}$.
- *Known initial state*: an initial state in K .
- *Known goal state*: a goal state in K .

- *Edge*: an $\langle s, s', o \rangle \in S \times S \times (F \cup B)$ triple where if $o \in F$, then $s' \in o(s)$, and if $o \in B$, then $s \in o(s')$.
- *Path*: an ordered sequence of edges in the form

$$\langle s_1, s_2, o_1 \rangle, \langle s_2, s_3, o_2 \rangle, \dots, \langle s_{k-1}, s_k, o_{k-1} \rangle,$$

where $k \geq 2$.

General objective: determine a path from s_0 to s^* , where s_0 is an initial state, and s^* is a goal state.

B. A Problem

One of the first applications developed in Smart Campus is a special service processing calendars that contain some events marked by students as important. If some of these events conflict in time with each other, the calendar service may suggest another schedule by replacing the time intervals of some events with other possible intervals. Suppose the students designates a schedule containing k events (E_1, \dots, E_k) , each with an initial time interval (from $T_{1,1}, \dots, T_{1,N_1}$ to $T_{k,1}, \dots, T_{k,N_k}$). The problem, which can be solved with the use of EBFS, is to determine a schedule of the same events such that no two time intervals are in conflict.

1) *State Space*: In this state-space representation, a state of the problem is represented by a k -tuple, the elements of which describe the currently set time intervals of each event. The IK states are arbitrarily chosen by the user and may contain interference between the time intervals of the events. Our goal is to eliminate this interference. In this model, initial states have no significance.

$$E_1 = \{T_{1,1}, \dots, T_{1,N_1}\}, \dots, E_k = \{T_{k,1}, \dots, T_{k,N_k}\}$$

$$S = \{\langle t_1, \dots, t_k \rangle \in E_1 \times \dots \times E_k : t_j \in E_j\}$$

$$K = \{\langle t_1, \dots, t_k \rangle \subseteq E_1 \times \dots \times E_k$$

$$\text{initial}(\langle t_1, \dots, t_k \rangle) = \text{true}$$

$$\text{goal}(\langle t_1, \dots, t_k \rangle) = \begin{cases} \text{true} & \text{if } \forall p \forall q (p \neq q \rightarrow t_p \cap t_q = \emptyset) \\ \text{false} & \text{otherwise} \end{cases}$$

2) Operators Over the State Space:

$$F = \left\{ \text{update}(l, r) \in (2^S)^S \right\},$$

$$l \in \{1, \dots, k\}, \quad r \in \{1, \dots, N_l\}$$

$$\text{update}_{l,r}(\langle t_1, \dots, t_k \rangle) = \langle t'_1, \dots, t'_k \rangle$$

$$\text{where } t'_j = \begin{cases} T_{l,r} & \text{if } j = l \\ t_j & \text{otherwise} \end{cases}$$

$$B = F$$

C. The EBFS Algorithm

The EBFS algorithm extends the BFS algorithm with the ability to run more than one breadth-first search starting from more than one state (the initially known states).

The EBFS algorithm stores a subgraph of the representation graph during the search. The main difference from BFS at this point is that in case of EBFS, the relationship between the nodes and each IK state is stored.

The full pseudocode of the EBFS algorithm can be found in [4]. The database of the algorithm stores for each node the state represented by the node as usual, the forward and backward status (open, closed, or not relevant), forward and backward parents, forward and backward children of the node, as well as the distance from and to each of the IK states.

VII. CONCLUSION AND FUTURE WORK

In this paper we introduced the intelligent services and outlined that the development of smart applications is a never-ending process. The underlying service architecture are now in a testing phase and several end-user application prepared but several more need to be created before we could call our system smart. The architecture fits well into the more general publish/subscribe based architecture of Smart City and Smart Campus applications as its extensible with new data sources providing the capability of integration of heterogeneous data. In the future, development of the Analytics module is a major goal since providing good analysis of the collected data can add more value to the services.

ACKNOWLEDGEMENTS

The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

REFERENCES

- [1] Attila Adamkó and Lajos Kollár. Extensible data management architecture for smart campus applications—a crowdsourcing based solution. In *WEBIST (1)*, pages 226–232, 2014.
- [2] Y. Atif and S. Mathew. A social web of things approach to a smart campus model. In *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCoM), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*, pages 349–354, Aug 2013.
- [3] Mikel Emaldi, Oscar Pena, Jon Lazaro, Diego Lopez-de ipina, Sacha Vanhecke, and Erik Mannens. To trust, or not to trust: highlighting the need for data provenance in mobile apps for smart cities. In *International Workshop on Semantic Sensor Networks, Proceedings*, pages 1–4, 2013.
- [4] Tamás Kádek and János Pánovics. Extended breadth-first search algorithm. *International Journal of Computer Science Issues*, 10(6):78–82, 2014.
- [5] Tamás Kádek and János Pánovics. Some improvements of the extended breadth-first search algorithm. *Studia Universitatis Babeş-Bolyai, Informatica*, 59(Special Issue 1):165–173, 2014.
- [6] Haim Kaplan, Ilia Lotosh, Tova Milo, and Slava Novgorodov. Answering planning queries with the crowd. *Proc. VLDB Endow.*, 6(9):697–708, July 2013.
- [7] Ingo Lütkebohle. The Apps for Smart Cities Manifesto. <http://www.appsforsmartcities.com/?q=manifesto>, 2012. [Online; accessed 15-December-2014].
- [8] Adam Marcus, David Karger, Samuel Madden, Robert Miller, and Sewoong Oh. Counting with the crowd. *Proc. VLDB Endow.*, 6(2):109–120, December 2012.
- [9] P. Saint-Andre. RFC 6120: Extensible Messaging and Presence Protocol (XMPP): Core., March 2011.
- [10] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, May 2006.
- [11] Róbert Szabó, Károly Farkas, Márton Ispány, András A. Benczúr, Norbert Bátfai, Péter Jeszenszky, Sándor Lakó, Anikó Vágner, Lajos Kollár, Csaba Sidló, Renátó Besenczi, Máté Smajda, Gergely Kövér, Tamás Szincsák, Tamás Kádek, Márk Kósa, Attila Adamkó, Imre Lendák, Bernát Wiandt, Timon Tomás, Ádám Nagy, and Gábor Fehér. Framework for smart city applications based on participatory sensing. In *Proceedings of the 4th IEEE International Conference on Cognitive Infocommunications*, pages 295–300, Dec 2013.

Machine-Learning

An overview of optimization techniques

Pedro Oliveira, Filipe Portela, Manuel Filipe Santos, António Abelha, José Machado.

Abstract— In an intelligent system the tasks roles is an essential play between learning and optimization. The Machine Learning is used to address a specific problem. However, the optimization of these systems are particularly difficult to apply due to the dynamic, complex and multidisciplinary nature. Nowadays we notice a constant research and development of new algorithms capable of extracting knowledge treated large volumes of data, thus obtaining better predictive results than current algorithms. There emerges ~~and~~ a large group of techniques and models that are best suited to the nature and complexity of the problem. It is in this regard that incorporates this work. The aim of this work is to present an overview of the most recent and most used optimization techniques in machine learning.

Keywords—Machine Learning, Optimization techniques, Literature review.

I. INTRODUCTION

THE Machine-Learning systems conception is to find patterns and realize automatic tasks recurring data to generalize pretended cases.

Machine-Learning (ML), can help discovering patterns and to perform certain tasks through the generalization of cases and the use of data. As the basis of these decisions are the learning and knowledge systems. These systems are enriched with information in the form of structured or unstructured data to better search, match and get the best forecasts and analysis of the problem in question. This issue raises fundamental philosophical questions about what constitutes "learning" in general, typically defined as: gain knowledge or skills, to study or experience; commit to memory; be warned, be informed; becoming aware; the behavior modification through interaction with the environment reasoning premises to conclusions. We can define information as data plus meaning (events) with significance, as knowledge plus experience can be considered wisdom in understanding the information [1].

This work was FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

Pedro Oliveira is with Information System Department, University of Minho, Portugal.

Filipe Portela is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (Corresponding author to provide phone: +351253510319; fax: +351253510300; e-mail: cfp@dsi.uminho.pt).

Manuel Filipe Santos, is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (e-mail: mfs@dsi.uminho.pt).

António Abelha and José Machado are with Algoritmi Research Centre, University of Minho, Braga, Portugal. (e-mail: {Abelha,jmac}@di.uminho.pt).

Its implementation is considered feasible and low cost compared to manual programming. Thus, as new data emerge, the more ambitious problems can be solved using the ML. As a result, they are widely used in computer science among others, such as web searches, spam filters, recommendations systems, ad placement, assignment rankings and fraud detection among many others [2].

In this part, learning processes, adaptation and optimization are explored through the use of algorithmic approaches [3]. These approaches are an attempt to extract rules / standards in the available data (typically using statistical techniques or data mining) [4], where the results are probabilities rather than certainties [1].

With reference to the above mentioned it is possible to observe that the optimization is part of ML. Most machine learning problems are reduced to optimization problems. Whereas the action of the ML analysis and solving problems of a specific set of data. The decision maker formulates the problem of selecting appropriate models families, transforming the data into a suitable format [5]. This type of model is typically trained to solve optimization problems of nuclear systems, which optimize the variables or parameters regarding the used function model. The computational mathematics research area intercepts with the level of nuclear optimization problems, predisposing theories and definitions that are an optimal solution based on ideal conditions. In order to evaluate the performance of optimization models in ML it was made a literature review as a tool to evaluate best practices / optimization algorithms. The result of applying this instrument should be able to bring competitive advantages to the ML. This paper presents an overview of the most significant techniques found in a deep literature review realized. The paper is divided in five sections, its essential includes an introduction, the considerations taken to make the review, the overview of Machine Learning and Metaheuristics and finally a brief conclusion.

II. LITERATURE REVIEW STRATEGY

This literature review was based on the research of concepts related to ML and optimization techniques. They were used several scientific research engines ScienceDirect; Web of Knowledge; Springer; IEEE Xplore; Google Scholar; B-on; Scopus. The choice of these articles followed preferred criteria such as: Date (preferably later articles to 2000) and / or

Relevance (preferably more than 10 citations per last year since the publication); Author (more publications on the keywords addressed).

III. MACHINE-LEARNING

ML is focused on developing systems that learn from the data [2] [6]. This involves a training phase where the system learns to complete certain tasks (predictive or classification) using a given data set containing information representative of the problem. After the training phase, the system is able to analyze new data having the same set of parameters and suggest a prediction. Unfortunately, there is no perfect method that is able to solve a particular problem, as there are several that offer best hits and forecasts easily [7] being dependent on the study area. This is an aspect that should be considered before developing a system based on these models and we will review.

A. Logistic regression

The Logistic regression (LR) [8] seeks to achieve the influence of independent variables in predicting categorically the dependence of a variable (which has a number of limit values). This technique is commonly useful for identifying in a dataset the most discriminating variables and its output can only assume predefined values (ex. Positive or negative). These models tend to be less robust than the Artificial Neural Networks (ANN) and Support Vector Machine (SVM) when we are dealing with a complex set of data. However, they are used simple linear models to process quick decisions as it is easier to interpret the output and how the decision was made [7].

B. Artificial neural networks

This mathematical model, known as artificial neural networks (ANN), is conceptually similar to SVM [7], interpret the learning process in the human brain using artificial neurons interconnected in a network that identifies patterns in data [9]. A neural network has some inputs and produces one or more outputs applying incremental learning algorithms to process and modify the intensity of the links between inputs, outputs and hidden layers of the network, with observed patterns among the data [7]. The adoption of neural models has several advantages. They are implemented without much statistical training, are endowed with skills which implicitly detect nonlinear relationships between complex dependent and independent variables and the ability to detect all possible interactions between predictor variables [7]. The disadvantages focus on rational behavior. The perception and the decision is implemented through the hidden layers which is trivial for the user to realize what was decided and why, which makes not prone to possible adjustments (because the model describes the error and the random noise rather than the underlying relationship the data) [10]. However, there have been efforts in the perception of this limitation [11].

C. Support vector machines

The Support Vector Machine (SVM), presented by Vapnik [12], are powerful and complex instruments that fit particularly

when the classification task is difficult [13]. Examples of an SVM model is a set of data points in space as to become divided into different categories for the widest possible space [7]. These instances are mapped getting divided with regard to its category, space and forecasting using the kernel trick [13]. It is an efficient method for problem solving in pattern recognition and regression and the analysis of handwritten documents, images and time series forecasting. [12].

IV. METAHEURISTICS

The technical meta-heuristics will be successful ~~in~~ when a given optimization problem achieving provide a balance between diversification and intensification. The intensification is needed in the search for parts in space with high quality solutions, and it is important in finding some promising areas on the accumulated research experience. The main differences between the existing metaheuristics are related to the way of achieving this balance [14]. The classification criteria can be used for the meta-heuristics, in terms of the features that follow in the research, memory feature, type of neighbor holding used or the number of current solutions made from one iteration to the next.

For a more formal classification [14], it is performed a meta-heuristics differentiation between Single-solution based and Population-based. In general, the single-solution based are more targeted towards enhancing, while the Population-based are oriented to the exploitation [15]. The main algorithms belonging to these categorizations are briefly discussed below.

A. Single solution based

Presented as meta-heuristics based on unique solution, also known as trajectory methods. They start with an initial solution and describe the trajectory in space research when moves away from that solution. Some may be considered as "smart extensions" local search algorithms. These methods include mainly simulated annealing, tabu search and others variants [16].

B. Population based

Population-based Metaheuristics handle a set (population) solutions instead of an initial solution. Most studies based on these methods are related to Evolutionary Computation (EC) inspired by Darwin's theory, where the population of individuals is modified by recombination and mutant operators, and Swarm Intelligence (SI), where the idea is to create computational intelligence to explore simple analogies of social interactions rather than purely individual cognitive abilities [15]. Variants of these issues will be addressed in the following subsections.

1) Evolutionary computation

Evolutionary Computation (EC), inspired by the ability of living things to evolve and adapt to their environment, based on the principles of Darwin. EC is the general term for several optimization algorithms. Usually associated with the term Evolutionary Algorithms (EA), EA are methods such as genetic algorithms [17], evolutionary strategies [18] Evolutionary

programming [19], genetic programming [20], differential evolution, among others, where there is a sharing in the form as the simulation of the evolution structure their ideas through selection processes, recombination and mutation breeding in order to develop better solutions. Briefly this class of algorithms [21] contains: Representation (definition of individuals); evaluative function; Population; the parent selection mechanism; variation operators, recombination and mutation; Survival Mechanisms (replacement). Afterwards, it is presented a set of algorithms that highlighted and emerged over the last years.

a) *Evolution strategy*

Evolutionary Strategy (ES) mimics the principles of natural evolution as a method for solving optimization problems.

Introduced by Rechenberg [22] and developed by Schwefel [23], the first ES algorithm was used to optimize experimental parameters. However, it is based on a population formed by a single progenitor through mutation which produces a single Gaussian downward. The selection criterion determines the ability of the individual in the intuited to become the progenitor of the next generation. Rechenberg proposed EE multimembered, introducing the concept of population, where more than one parent may jointly generate a single downward. With this, you can have additional recombination operations, when two parents chosen randomly recombine to give a child, subject to change. The selection process now takes into account the worst extinction, which can be both a parent and a child, in order to maintain constant population size. Mutation is accomplished by numbers distributed with zero mean and standard deviation (determines the size of the mutation) and is easy to understand that the parameters of the distribution compromise the performance of the search algorithm. The simplest way is to specify the changing mechanism to maintain its constant over time.

There are several approaches to this method, however, recently it was introduced a method by Hansen et al. [18] Covariance Matrix titled Adaptation Evolution Strategy (CMA-ES). It proved to be very effective and it is currently the most used in the range of evolutionary algorithms for local optimizations as well as for global optimizations [24].

b) *Differential evolution*

One of the most popular algorithms for continuous optimization problems is the Differential Evolution (DE). Proposed by Storn and Price [25] in order to solve a polynomial fit problem, proved to be a very reliable optimization strategy for other tasks.

As with any EA, a population of candidate solutions is randomly selected for a particular optimization task. In each process of evolutionary generation, new individuals are created by applying operators (crossover and mutation). The ability of the resulting solutions are evaluated by each individual of the population against a young guy (mutant), where it is created by recombining the individual of the population with another individual created by mutation, in order to determine which one will be maintained for the next generation [15]. The main

advantage of DE is that they have less control parameters (only three entries), which control the search process (population size, differentiation and crossing). Consequently, these parameters are fixed, which does not become trivial to set priorities in the parameters by a certain problem. Thus, some authors have developed strategies in setting parameters according to experience learning [15].

DE today is one of the most popular heuristics to solve single-objective optimization problems in continuous search spaces, where its use has been expanded to multi-objective problems. However, there are gaps in slow convergence and stagnation of the population. More variants, details and applications are referred to articles like Neri and Tirronen [26].

2) *Swarm Intelligence*

Swarm Intelligence (SI) is a paradigm of distributed intelligence and innovative in optimizing troubleshooting inspired by collective behavior of many living beings. Typically comprise a population of agents (able to perform various tasks) interacting among themselves and with the surrounding environment. The absence of a single control structure, local interactions among these agents lead to the emergence of self-organizing global behaviors [15].

Many optimization algorithms such as Ant colony optimization, Particle Swarm Optimization, Bacterial foraging optimization, Bee Colony Optimization, Artificial Immune Systems, Firefly algorithm, Gravitational search, Biogeography-Based Optimization, Bat algorithm and Krill herd are inspired by the metaphors of this behavior [27]. The following subsections examine in general some of these new algorithms paradigms.

a) *Ant colony optimization*

Ant colony optimization (ACO) is a meta-heuristic inspired by the behavior of real ants in search of food for solving combinatorial optimization problems introduced and surveyed by M. Dorigo [28]. When looking for food the ants begin by analyzing the area around their nest. Then along the trajectory releases a track with chemical pheromone on the ground in order to schedule a favorable path to guide other ants to the discovered source of food [28]. After that, the shortest path between the nests is labeled with a higher concentration of pheromones which in turn attracts more ants. With this, it is expected to explore the characteristics of ant colonies to build solutions with the exchange of information on the quality and the communication scheme for optimization problems.

ACO algorithms have different proposes but all share the same features. Their discussion, research and applications can be found at many research articles [29] where the authors relate ACO with other variants. More recently Angus and Woodward [30] argued that these algorithms will be a great advantage, and common, when they are systematically applied in real-world applications with variable data in terms of time and availability.

b) *Bat algorithm*

The bats are the only mammals with wings that have at least 1000 different species that represents up to 20% of all mammal

species. Bat algorithm (BA), developed by Xin-She Yang in 2010 [31], represent a particular bat specie behavior, microbat, that emit sound pulses and listen to the echo from the surrounding objects, called *echolocation*. They use this short frequency-modulated sound pulses to sense distance and orientation of the target, type of prey and their moving speed in the dark. This characteristic has many advantages, for example it can provide very quick convergence by switching from diversification to intensification. Praising the advantages, it can summarize the key points in Frequency tuning, automatic zooming and parameter control. From this, many other methods and strategies have been attempted to increase the diversity of the solution and to enhance performance. With this, at least nine variants were emerged to explore this differences. Concluding this relevance over the years, BA is easy to implement and can solve a wide range of problems. On a particular comparison case obtained from Khan and Sahai [32]. Classification problems and an eLearning case showed that BA recurred less functions evaluations to reach optimal solution with lower average error facing other techniques like PSO or GA.

c) *Bee colony*

Bee colony optimization algorithm-based (BCOB) are a new generation of algorithms inspired by the behavior of bee colonies. They have resources that can be used as models for SI and collective behavior as waggle dance (communication), foraging, queen, task selection bee, collective decision-making, the mating nest site selection in flight, marriage settlements, flowering and navigation systems [33]. With this, several algorithms based on these behaviors have been proposed in order to replicate their knowledge. A literature review on algorithms inspired by the behavior of bees in nature and its applications can be found at Karaboga and Akay article in [33].

d) *Bio-geography*

Developed by Dan Simon in 2008 [34], the Biogeography-based optimization algorithm (BBO) was influenced by biogeographical balance islands [35], which deals with the change of balance between immigration of new species and the emigration of species already installed. Each island is a set of candidate solutions, with a particular index Suitable variable (VS) and the other for the evaluator titled habitat suitability index (HSI) is used to measure the efficiency and effectiveness of the solution. In this algorithm, each individual has its own rate of immigration and emigration, and good solutions (islands with many species) tend to share their resources with weak solutions (islands with few species). Poor solutions are receptive to new species of good solutions [35].

There are other important factors that influence the migration rates between habitats, such as distance to the neighboring habitat, its size, climate (rainfall and temperature), vegetation, animal diversity and human activity that have not been considered. Thus, Haiping Ma [36] explored six different types of migration, and tested its performance with wide ranges and dimensions through 23 benchmark functions. The results showed significant positive changes in performance compared to linear models in most benchmarks.

e) *Firefly algorithm*

In countries like Portugal, in the summer people are fascinated with the light of the fireflies. Xin-She Yang adapt this behavior to inspire a development of a metaheuristic algorithm called Firefly algorithm. The production of short and rhythmic flashes offer a unique pattern of this species, until now only three behaviors were interpreted in their communication, hunt skills and protection [37]. A simple idealization of the firefly algorithm structure can be realized in three points: the fireflies will be attracted to others regardless their sex; light brightness is proportional to attractiveness and their search is random; the bright is determined by the landscape of the objective function. After this, swarming agents can interact with others providing mechanisms of intensity, but it can also offer some diversification based by the series of Brownian motion that obeys a Gaussian Diffusion or a non-Gaussian diffusion, whereas the Gaussian diffusion showed more improvements than the others [31].

In the last years, the standard FA appeared to be efficient, however, other variations, or some modifications expanded quickly and it is impossible to list all the variants, though some of them can be found at Yang [31].

The relevance of this algorithm was widely discussed because of its multi-modal characteristics, the capability to handle the problems efficiently, with a fast convergence rate in general, global and local search problems to every problem domains (nature-inspired optimization algorithm).

Applications with this method are presented, for example, by Banati et al. [38] with a hybridized FA concerned on preprocessing techniques in machine learning. Recurring at four different medical datasets, purposing a simulation of the attraction systems of real fireflies that find the best feature selection procedure. This method beats others features selections in terms of time and optimality [39].

f) *Gravitational search*

Gravitational Search Algorithm (GSA) introduced by Rashedi et al. [40], presented a construction of a method based on the law of gravity and the notion of mass interactions. Using the theory of Newton, it can considerate each mass a solution, and the algorithm navigate adjusting the gravitation and inertia masses. Over the time, the masses will be attracted by the heaviest one, presenting an optimum solution in the space research. This can be considered as an isolated systems obeying the laws of gravity and motion.

Understanding this laws it is possible to interpret this algorithm in some relevant points: each agent can observe the others through the gravitational force; this force acts in the neighborhood of the agents, providing the capability to see his space around; Agents with greater gravitational mass have higher performance, pointing to the best agent; the adaptive learning rate is related with the agents that have heavy inertia mass, turning their moves slowly and the search space more reduced; it is a memory-less algorithm with fast convergence.

In the last six years, the GSA algorithm had been used to derive in other variants, creating at least twenty new types of them. With this importance was necessary a comparison to

others techniques, performed using datasets like Iris, wine, glass and cancer to classify the accuracy and rank. In almost all of them the GSA provide best result among the other techniques [41].

g) *Krill Herd*

The now Krill herd presented by Gandomi and Alavi [42], was inspired from the krill herding motions to solve optimization problems. This motions are determined by three essential actions of time-dependent position: reaction in the presence of others, searching for provisions that contains a global and local optimizer parallelization, and their diffusion behavior for the adaptive evolutionary operators (mutation and crossover). This exempt the derivation of information, because the use of stochastic random search.

A particular part and a great advantage of this algorithm related to other nature-inspired algorithms is the fact that only time interval should be fine-tuned for each problem.

Characteristics of each agent can contribute to the moving process according to its fitness, their neighbor attract/repulse the individual, acting as a local search and the global best is regarded according to the center of food of all the krill individuals.

Meanwhile to prove the efficiency of the proposed algorithm four different KH algorithms were derived and created: KH without any genetic operators, KH with crossover operator, KH with mutator operator and with both. After this each one was tested for solving benchmark problems, and it was concluded that KH without any genetic or with crossover and mutator operators showed better results than many others algorithms [42].

Applications of this methods are scarce because of it is relatively recent presentation. However Wang et al. [43] proposed a hybrid krill herd algorithm facing with eight other population-based optimizations methods throw mathematical functions. This benchmark functions indicates hybrid KH algorithm like the more powerful and efficient optimization algorithm of population-based problems [43].

h) *Particle swarm optimization*

Presented in 1995 by Kennedy and Eberhart [44], the particle swarm optimization (PSO) is entitled as a global optimization technique that uses metaphors behavior in groups of birds when they are flying to abroad optimization problems. There are some differences between PSO and evolutionary optimization which were exposed and discussed in the paper [45]. In this algorithm, autonomous entities (particles) are randomly generating events in space research, where each entity is a candidate solution to the problem at hand. A cluster consists of a number of particles around a certain dimensional space research, where there is some type of topology [15], represented by a location and velocity, writing the interconnections between the particles memorizing the previous best position.

Kennedy et al. [46] concluded that this tends to converge topology for the likelihood of getting stuck in local optima, however, this topology is slower but explores more deeply and usually ends in the best optimum. It has been implemented a lot

of effort in understanding the functioning of the EPO algorithm in analyzing the trajectories [47] and why fail under certain conditions. The EPO formulation in parallel implementation was also discussed by [48] and how to adapt to this type of optimization.

C. *Swarm Intelligence Analysis*

Table 1 presents an analyses of the metaheuristics presented. A set of features are measured and assessed through their functionality: H – High; M – Medium; L – Low; N – No; Y – Yes; C – Crossover; Mu – Mutators – Selector.

Table 1 – Analysis of Swarm Intelligence Techniques

	PSO	KH	BBO	GSA	BA	FA	BCOB	ACO
Speed training	L	H	H	H	H	H	M	L
Memory Usage	H	L	L	L	L	L	M	M
Predictive accuracy	H	H	M	M	M	H	M	M
Interpretability	L	M	L	M	M	M	H	H
Predicting speed	L	H	M	H	H	H	H	M
Fitting speed	M	H	M	H	H	H	M	L
Handle categorical predictors	N	Y	N	N	Y	Y	N	N
Parameter adjust	Y	Y	Y	N	Y	Y	N	N
Genetic operators	S	C; Mu	Mu	C; Mu	Mu; S	Mu; S	N	N
Exploitation (local)	H	H	L	L	H	H	L	L
Exploration (global)	L	H	H	H	H	H	H	H

V. CONCLUSION

This paper presented briefly a wide range of perspective in what was pioneer and what is now in matters of learning and optimization. The vision created offers a new panorama in solving old and new problems, single or population based, that concludes a necessity for looking sharper, septic and adopt the potential of this new SI techniques. Excepting the CMA-ES, Cuckoo Search or hybrid variations [31] that was not taken in this review, the nature inspired algorithms take almost all the best results in various forms of benchmarking and applications, combining advantages in terms of classification criteria. With this effort, the scientific community has a guideline about which are the most used optimization algorithms in ML to single and population based problems. At same time this overview offers a set of papers (references) that can be consulted in order to make a deeper analysis of each algorithm.

ACKNOWLEDGMENT

This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

REFERENCES

- [1] Fulcher, J., & Jain C., L. (2008). Computational Intelligence: A Compendium. W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
- [2] Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78. Retrieved from <http://www.mendeley.com/catalog/few-useful-things-know-about-machine-learning/>.
- [3] Pardalos, P. M., & Romeijn, H. E. (2005). Handbook of Global Optimization vol.2. Retrieved December 09, 2014, from http://www.optimization-online.org/DB_FILE/2002/03/456.pdf
- [4] Shearer C, Caron P (2002) Handbook of Data Mining and Knowledge Discovery. Oxford University Press, UK.
- [5] P. Bennet, K., & Parrado-Hernández, E. (2006). The Interplay of Optimization and Machine Learning Research. *Journal of Machine Learning Research*, 7. Retrieved from <http://www.jmlr.org/papers/volume7/MLOPT-intro06a/MLOPT-intro06a.pdf>.
- [6] Rogers S, Girolami M. A first course in machine learning. New York, NY, USA: CRC Press Inc.; 2011. (<http://books.google.co.uk/books?id=rdQ1daD8BH8C>).
- [7] Fraccaro, P, et al. Behind the screens: Clinical decision support methodologies – A review. *Health Policy and Technology*, (2014). doi:10.1016/j.hlpt.2014.10.001
- [8] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York, NY, USA: Springer; 2009.
- [9] Picton P. Neural networks. New York, NY, USA: Palgrave Macmillan; 2000. (<http://books.google.it/books?id=mBk6qAAACAj>).
- [10] Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol*. Elsevier 1996;49(11):1225–31. November.
- [11] Rudy Setiono, Wee K. Leow, and James Y. L. Thong. Opening the neural network black box: an algorithm for extracting rules from function approximating artificial neural networks. In *ICIS '00: Proceedings of the twenty first international conference on Information systems*, pp. 176–186, Atlanta, GA, USA, 2000. Association for Information Systems.
- [12] V. Vapnik, *Statistical Learning Theory*, Wiley Press, New York (1998) 493–520.
- [13] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other Kernel-based learning methods. New York, NY, USA: Cambridge University Press; 2000. Available from: (<http://books.google.it/books?id=B-Y88GdO1yYC>).
- [14] M. Birattari, L. Paquete, T. Stützle, K. Varrentrapp, Classification of Metaheuristics and Design of Experiments for the Analysis of Components, Technical Report AIDA-01-05, FG Intellektik, FB Informatik, Technische Universität Darmstadt, Darmstadt, Germany, 2001.
- [15] Boussaïd, I., Lepagnot, J., & Siarry, P. (2013). A survey on optimization metaheuristics. *Information Sciences*, 237, 82–117. doi:10.1016/j.ins.2013.02.041.
- [16] E.G. Talbi, *Metaheuristics: From Design to Implementation*, first ed., Wiley-Blackwell, 2009.
- [17] D. Beasley, D. Bull, R.R. Martin, An overview of genetic algorithms. Part i: fundamentals, *University Computing* 15 (1993) 58–69.
- [18] N. Hansen, A. Ostermeier, A. Gawelczyk, On the adaptation of arbitrary normal mutation distributions in evolution strategies: the generating set adaptation, in: *Proceedings of the 6th International Conference on Genetic Algorithms*, Morgan Kaufman Publishers Inc., San Francisco, CA, USA, 1995, pp. 57–64.
- [19] D.B. Fogel, *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*, IEEE Press, Piscataway, NJ, USA, 1995.
- [20] Y. Shan, R.I. McKay, D. Essam, H.A. Abbass, A survey of probabilistic model building genetic programming, in: *Scalable Optimization via Probabilistic Modeling*, Studies in Computational Intelligence, vol. 33, Springer, 2006, pp. 121–160.
- [21] L. Bianchi, M. Dorigo, L.M. Gambardella, W.J. Gutjahr, A survey on metaheuristics for stochastic combinatorial optimization, *Natural Computing* 8(2009) 239–287.
- [22] I. Rechenberg, *Cybernetic Solution Path of an Experimental Problem*, Technical Report, Royal Air Force Establishment, 1965.
- [23] Beyer, H.-G., Beyer, H.-G., Schwefel, H.-P., & Schwefel, H.-P. (2002). Evolution strategies – A comprehensive introduction. *Natural Computing*, 1, 3 – 52. doi:10.1023/A:1015059928466.
- [24] N. Hansen, The CMA evolution strategy: a comparing review, in: J. Lozano, P. Larranaga, I. Inza, E. Bengoetxea (Eds.), *Towards a New Evolutionary Computation. Advances on Estimation of Distribution Algorithms*, Springer, 2006, pp. 75–102.
- [25] R.M. Storn, K.V. Price, Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces, *Journal of Global Optimization* 11 (1997) 341–359.
- [26] F. Neri, V. Tirronen, Recent advances in differential evolution: a survey and experimental analysis, *Artificial Intelligence Review* 33 (2010) 61–106
- [27] A. Engelbrecht, *Fundamentals of Computational Swarm Intelligence*, Wiley, 2006.
- [28] M. Dorigo, C. Blum, Ant colony optimization theory: a survey, *Theoretical Computer Science* 344 (2005).
- [29] M. Dorigo, T. Stützle, The ant colony optimization metaheuristic: algorithms, applications, and advances, in: F. Glover, G. Kochenberger (Eds.), *Handbook of Metaheuristics*, International Series in Operations Research & Management Science, vol. 57, Springer, New York, 2003, pp. 250–285
- [30] D. Angus, C. Woodward, Multiple objective ant colony optimization, *Swarm Intelligence* 3 (2009) 69–85.
- [31] Xin-She Yang, *Nature-Inspired Optimization Algorithms*, School of Science and Technology, Middlesex University London, Elsevier, 2014.
- [32] K. Khan, A. Sahai, A comparison of BS, GS, PSO, BP and LM, for Training Feed forward Neural Networks in e-Learning Context, *MECS Press*, 2012, 7, 23-29.
- [33] D. Karaboga, B. Akay, A survey: algorithms simulating bee swarm intelligence, *Artificial Intelligence Review* 31 (2009) 61–85.
- [34] D. Simon, Biogeography-based optimization, *IEEE Transactions on Evolutionary Computation* 12 (2008) 702–713.
- [35] R. MacArthur, E. Wilson, *The Theory of Biogeography*, Princeton University Press, Princeton, NJ, 1967.
- [36] H. Ma, An analysis of the equilibrium of migration models for biogeography-based optimization, *Information Sciences* 180 (2010) 3444–3464.
- [37] Lewis SM, Cratsley CK. Flash signal evolution, mate choice and predation in fireflies. *Ann Rev Entomol* 2008;53(2):293–321
- [38] H. Banati, M. Bajaj, Firey based feature selection approach, *IJCSI International Journal of Computer Science Issues* 8 (4) (2011) 473-480.
- [39] Fister, I., Yang, X. S., & Brest, J. (2013). A comprehensive review of firefly algorithms. *Swarm and Evolutionary Computation*, 13, 34–46. doi:10.1016/j.swevo.2013.06.001
- [40] Rashedi E, Hossein Nezamabadi-Pour H, Saryazdi S. GSA: a gravitational search algorithm. *Inform Sci* 2009;179(13):2232–48..
- [41] Systems, I. J. I., & Sahoo, G. (2014). A Review on Gravitational Search Algorithm and its Applications to Data Clustering & Classification, (May), 79–93. doi:10.5815/ijisa.2014.06.09
- [42] Gandomi AH, Alavi AH. Krill herd: a new bio-inspired optimization algorithm. *Commun Nonlinear Sci Numer Simul* 2012;17(12)
- [43] Wang, G., Guo, L., Wang, H., Duan, H., Liu, L., & Li, J. (2014). Incorporating mutation scheme into krill herd algorithm for global numerical optimization. *Neural Computing and Applications*, 24, 853–871. doi:10.1007/s00521-012-1304-8
- [44] J. Kennedy, R. Eberhart, Particle swarm optimization, *IEEE International Conference on Neural Networks* 4 (1995) 1942–1948.
- [45] P. Angeline, Evolutionary optimization versus particle swarm optimization: philosophy and performance differences, in: V. Porto, N. Saravanan, D. Waagen, A. Eiben (Eds.), *Evolutionary Programming VII, Lecture Notes in Computer Science*, vol. 1447, Springer, Berlin, Heidelberg, 1998, pp. 601–610.
- [46] J. Kennedy, R. Eberhart, Y. Shi, *Swarm Intelligence*, Morgan Kaufman, San Francisco, 2001.
- [47] M. Clerc, J. Kennedy, The particle swarm – explosion, stability, and convergence in a multidimensional complex space, *IEEE Transactions on Evolutionary Computation* 6 (2002) 58–73.
- [48] A. Banks, J. Vincent, C. Anyakoha, A review of particle swarm optimization. Part i: background and development, *Natural Computing* 6 (2007) 467–484. doi: 10.1007/s11047-007-9049-5.

The Personalized Recommendation Technology for Online Courses with Combinational Algorithm

Minjuan Wang, Jun Xiao, Bingqian Jiang, and Junli Li

Abstract—In the age of information explosion, it is important to improve the utilization rate of educational resources. This paper suggests a personalized recommendation system for online courses based on combination technology. The system mainly makes use of recommendation technology based on association rules, based on content, and based on collaborative filtering to implement the personalized recommendation for online courses and to complete the system evaluation and testing. The system, which proves to be able to improve the utilization rate of educational resources and promote the learning autonomy and efficiency of students, has obtained remarkable results in its application in Shanghai Lifelong Learning Network.

Keywords—Combinational algorithm, knowledge and data technology, intelligent learning systems, personalized recommendation

I. INTRODUCTION

With the rapid development of Mobile Internet and the progress of educational informationization, fundamental changes have occurred to teaching and learning, which spurred the widespread acceptance of online. In the meantime, the massive digital learning resources in the data era also enriched people's learning methods and experience. However, instead of quenching people's "thirst for resources", the massive learning resources in fact increase the burden on resource acquisition, and the actual resource needs are inundated by massive disorganized resources. Therefore, how to acquire the content of interest from the massive data becomes an issue concerned and studied by scholars and the academic circle [1]. However,

This paper is supported by "Shu Guang" award "MOOCs design and empirical research oriented Shanghai lifelong learning (13SG56)" from the Shanghai Municipal Education Commission and Shanghai Education Development Foundation. It is also supported by the 2014 Shanghai education scientific research key project "The Study of online learning mode for Shanghai lifelong learning (A1403)". Besides, thanks for the support of Science and Technology Commission of Shanghai Municipality research project "Shanghai Engineering Research Centre of Open Distance Education (13DZ2252200)".

Jun Xiao is with the Shanghai Engineering Research Centre of Open Distance Education, Shanghai Open University, Shanghai 200433 China. (phone: (+86)021 25653263; fax: (+86)021 25653263; email: xiaoj@shtvu.edu.cn).

Minjuan Wang is professor of Learning Design and Technology at San Diego State University; and distinguished research professor of Shanghai International Studies University. San Diego, USA. (email: mwang@mail.sdsu.edu)

Bingqian Jiang is with the Department of Education Information Technology, East China Normal University, Shanghai 200062 China. (email:51130104045@student.ecnu.edu.cn).

in the educational sector, people mainly focus on how to acquire personalized, intelligent and adaptive learning resources in digital learning. One of the solutions to this problem is intelligent personalized recommendation system.

The intelligent personalized recommendation system is a recommendation technology that push resources to users, according to their personal preferences, such as interest, hobby, occupation or professional trait and so forth [2]. With the continuous acquisition of user information and behavior data, the recommendation quality of the intelligent recommendation system is on the track of constant self-improvement to reach the goal of precise recommendation. Taking into consideration the characteristics of big data era, we studied the intelligent learning recommendation technology in the field of online education, and implemented a personalized recommendation system for intelligent learning based on combinational algorithm.

II. LITERATURE REVIEW AND TECHNICAL THEORY

Personalized recommendation system first emerged in 1990s as a singular concept. In the U.S. Conference on Artificial Intelligence in 1995, universities such as Stanford and Carnegie Mellon [3] proposed the idea of personalized recommendation system. It is generally agreed that this conference marks the formal beginning of the era of personalized recommendation technology. Personalized recommendation system was first used in the e-commerce sector. With the development of digital learning resources and the growing popularity of the "learner-oriented" educational concept, personalized recommendation technology begins to be used in the educational sector.

In the digital learning system, personalized recommendation can commendably make up for the deficiencies existing in traditional resource recommendation method, meet various learning requirements of different learners, and promote users' loyalty to digital learning system by enhancing user experience. With the fast growing of digital learning resources, many researchers carry out studies on personalized recommendation for online learning from different perspectives, such as learning resource recommendations by means of Web log mining method; learning resource recommendations based on collaborative filtering technology; ontology learning resource recommendations based on user model; resource recommendation based on learner characteristics; and

personalized resource recommendation base on semantic network. The mainstream recommendation technology is categorized into three kinds, according to the differences in implementation algorithms and methods, namely, recommendation based on association rules, recommendation based on content filtering and recommendation based on collaborative filtering [4]. New hybrid recommendation algorithm can be generated by synthesizing the afore-mentioned three recommendation methods [5] [6].

A. Recommendations Based on Association Rules

Working principles for recommendations based on association rules: First, the administrator formulates a series of rule entries, then the inter-item association is evaluated based on the formulated rules, and the closely-associated items are pushed to the users. When making recommendations, the system will analyze the interest, hobby or access record of the current user, and then recommend the resource or items that might interest the user according to the rules formulated in advance. However, the personalized recommendation based on association rule is unable to generate rules or make dynamic changes, and it can only recommend to the users the resource entries fitted into their original interest according to the formulated rules, but unable to recommend other high-quality resources or discover the potential points of interest of the users [7].

B. Recommendations Based on Content Filtering

Content filtering recommendation technology is the most basic method of information filtering, and it is a recommendation technology proposed at a relatively earlier time [8]. The working principle of content filtering is: Such technologies as probability statistics and machine learning are adopted for filtering, where, a user interest vector is first used to represent the user's information request; then, segmentation, indexing, weighing of word frequency statistics, are carried out in the text collection to generate a text vector. Finally, the similarity between the user vector and text vector is calculated and the resource entries with high similarities are sent to registered users of the user model. The content filtering technology is applicable to the learning resources of text recommendation type instead of multi-media recommendation type.

C. Recommendations Based on Collaborative filtering

The concept of collaborative filtering can be traced back to the last century, and it was initially put forward by Goldberg, Oki, Nichols and Terry in 1992 and firstly applied in Tapestry system. As the first-generation product of collaborative filtering technology, Tapestry system is an unsound system, with many defects [9]. However, with the technological development, collaborative filter (CF) has become the personalized recommendation technology that is most studied and most widely applied today, and this technology is also the core algorithm for the system mainly used by this study.

Different from the afore-mentioned two recommendation technologies, collaborative filtering recommendation needs to generate user recommendation on the basis of analyzing

resource content and calculating the matching degree between the resources and the user, and the recommendation is generated based on the user's rating for the resources. The basic principle is as shown in Fig.1: First, the system calculates the similarity between giver users (or items), and then the nearest neighbor set of the target user (or item) is searched; finally, the user's rating for the target item is predicated according to the rating given by the user (or item) in the nearest neighbor set [10]-[13].

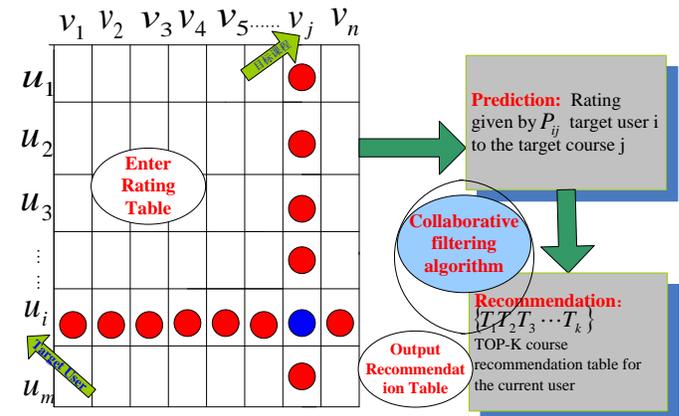


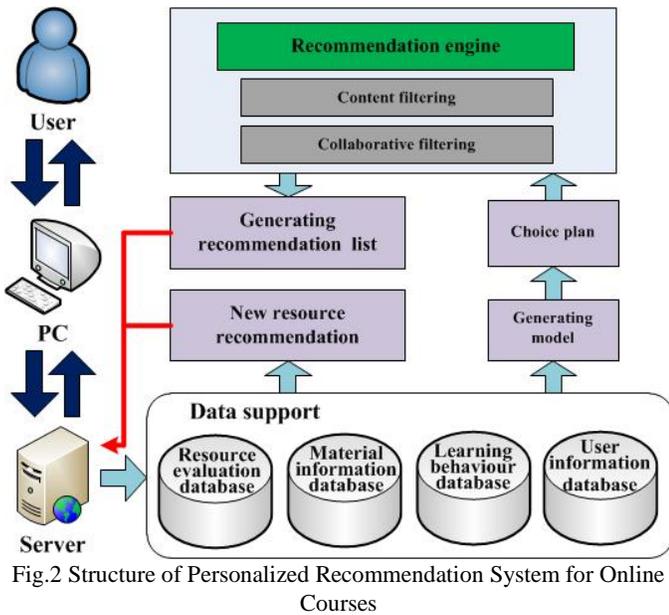
Fig. 1 Recommendation Process of Collaborative filtering System

The study mainly uses the collaborative filtering algorithm in conjunction with the combinational algorithm, developed by combining the recommendation technology based on association rules with the recommendation technology based on content filter, to implement the personalized recommendation system for online courses.

III. DESIGN OF PERSONALIZED RECOMMENDATION SYSTEM FOR ONLINE COURSES

A. Design of System Structure

As an integral part of lifelong learning system for Shanghai residents, Shanghai Lifelong Learning Network (www.shlll.net) is the main learning platform for online learning of Shanghai residents, and it provides the learners with a personalized user environment that enriches the user experience and constructs a personalized teaching and management system. Within half a year since the establishment of this Learning Network, the clicks on the learning platform have reached 2530,000. Users have grown to nearly 200,000 and the courses have exceeded 1,500. Facing extensive learning groups and massive course resources, Shanghai Lifelong Learning Network provides an excellent application environment and user data information for the design of personalized recommendation system for online courses. It has also become an important approach to further promote and enhance Shanghai Lifelong Learning Network to recommend to the learners the course resources that meet their learning requests. Therefore, based on Shanghai Lifelong Learning Network, this study adopts combinational algorithm to design the personalized recommendation system for Shanghai Lifelong Learning Network, as shown in Fig. 2.



The system model mainly includes three main functional modules: data support module, combinational algorithm recommendation engine module, and new source recommendation module [14].

B. Main Functional Modules of System

1) Data support module

Data support module refers to an information database, including four data tables: user information table, learning behavior data table, resource information table, and resource rating data table.

User information table: The personal information of users including the basic registration information and other relevant information obtained through Web data mining technology is stored in this table, such as interest, habit and resource preference, etc. To improve the precision of collaborative filtering recommendation, it is required that the system records user's personal information as specific as possible.

Learning behavior data table is used to preserve the learning behavior record of learners in the study process. The system, through tracking and recording various behavior data of learners, analyses and extracts the behavior data that can better reflect learners' resource preference (such as the downloading, reading, collection and recommendation of resources), and record it in the table. The learning behavior data are the data source of the implicit rating by users for resource entries.

Resource information library: Various learning resource information is saved in the system, such as courseware, cases, examination questions, news and literature, etc.

Resource scoring data table: It preserves the rating information for learning resources by learners. This table is the main data support for collaborative filtering algorithm. The collaborative filtering algorithm generates recommendations for users by analyzing the user-resource rating data, calculating the inter-user or inter-resource similarity.

2) Combinational Algorithm Recommendation Engine Module

The engine is the core module of recommendation system and also the centurms to implement personalized recommendation for learning resources, and the implementation process is shown as Fig.3.

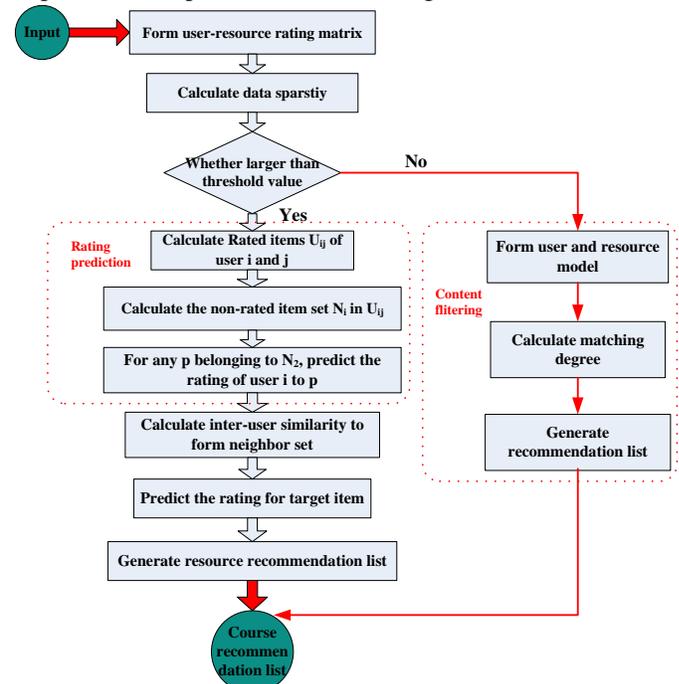


Fig. 3 Flow Chart for Combinational Algorithm of Personalized Video Resource Recommendation System

The algorithm process of the whole recommendation engine can be summarized into the following steps:

Step1: Retrieve the database and form the user-resource rating matrix;

Step 2: Calculate the data sparsity, which is defined as:

$$\text{Sparsity} = \frac{\text{User} - \text{Number of Categorized Entries of Courses}}{\text{Numer of Users} \times \text{Number of Resources}}$$

Step3: According to the sparsity degree, the method can be selected to correct the collaborative filtering algorithm. Here we set a threshold value "Th-value" as the critical value to select the evaluation prediction or content filtering. When $\text{Sparsity} < \text{Th-value}$, it is considered that the system is in the state of "cold boot" or "pre-cold boot". At this time, the content filtering shall be selected as the correction for collaborative filtering algorithm. When $\text{Sparsity} > \text{Th-value}$, the evaluation prediction algorithm shall be adopted for correction;

Step 4: form the adjacent users and generate the recommendation courses finally, so as to form the recommendation list in accordance with TOP-K.

3) New resource recommendation module

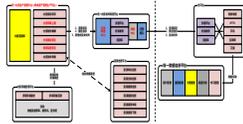
This module is mainly designed for the "cold boot" problem in collaborative filtering recommendation technology. Its main function is to analyze the interest, hobby and major (profession) category of each learners, and recommend the latest resources

in the relevant fields for them. However, if a newly added resource has not been accessed or rated by learners, it will never get the chance to be recommended by this system. By adding such module, the cold boot problem in collaborative filtering can be conquered to certain degrees, so as to improve the clicking rate of learning resources that have been newly added to the library [15]. In Shanghai Lifelong Learning Network, the effective access mechanism of the new courses is available.

C. System Evaluation and Test

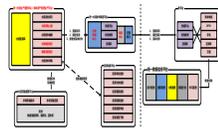
There are many standards to evaluate the effect of a recommendation system, and mean absolute error (MAE) is the simplest and most common performance evaluation standard in the recommendation system. The precision and recall evaluation criteria are often used in the evaluation for the TOP-N-based recommendation system [16], [17].

To evaluate the recommendation precision of the recommendation system, we need to divide the rating set into training set and test set. The MAE calculation is to obtain the mean absolute value of the differences between the actual value of all the evaluation data in the test set and the prediction value obtained by system calculation. For the target user u , supposing that the number of items that have been rated by the user is T_u , the corresponding collection of actual value is $\{q_1, q_2, q_3 \wedge q_{T_u}\}$, and the collection of predicative value obtained by the system is $\{p_1, p_2, p_3 \wedge p_{T_u}\}$, then, the mean absolute error of user u is MAE_u .



(1)

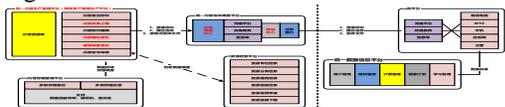
For the whole recommendation system, the MAE of all users is:



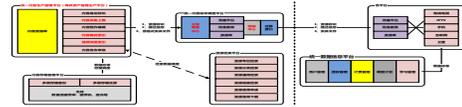
(2)

Where, M represents the number of users in the test set. Smaller value of MAE indicates higher accuracy of the predicative value obtained through system calculation and higher system recommendation quality.

Precision and recall are used to evaluate the system, and it is required to apply relevant algorithms in the training set to train system models so as to generate the top-N set of target users. Finally, the items appearing in the test set and Top-N set at the same time are added in the hit set. The recommendation precision to the target users refers to the probability of the items (N represents the number of items) recommended to the users appearing in the test set at the same time. The recommendation recall of the target users refers to the probability of the items already selected by the users from the test set appearing in the Top-N set for the users at the same time, as shown in the following two formulas:

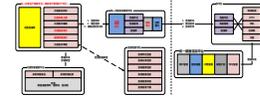


(3)



(4)

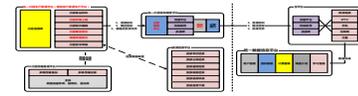
For the whole recommendation system, the recommendation precision and recall can be defined as the mean value of the recommendation precision and recall of all system users. In the item-based top-N recommendation system, the following method is usually adopted to calculate the system recall; an entry is randomly selected from the rating data of each user from the system and added into the test set, and the remaining rating data is used to train the system model and added into the training set. The model obtained through training can be used to calculate the top-N set for each user, and the recall can be calculated as follows:



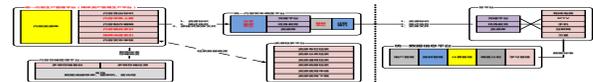
(5)

Where, m represents all users of the system. Generally, this measurement method is called hit-rate standard. When $recall=1$, it indicates that the system can recommend all the items wanted by users; however, when, $recall = 0$, it indicates the system can not recommend any item wanted by users.

From (3) and (4), the recommendation precision and recall are contradictory to some extent, i.e., the improvement of precision often leads to the reduction of recall. Therefore, to better evaluate the performance of the recommendation system, it usually requires to consider the two measurement methods in combination, i.e., the two are combined in a certain way to form a comprehensive evaluation index, and the common combination methods include F1 and E-measure:



(6)



(7)

The study draws references from the mainstream test method of recommendation engine, and evaluates and tests the system by recording the time of user click, login users, recommended course ID, original course ID and the sum of the number of times that the recommended courses appear, and testing the proportion between the times that the recommended courses appear and the number of clicks of the recommended courses, the proportion between the clicks of the recommended courses and the number of the selected courses. Through the system test, it is found that the engine can effectively detect the users' interest and, therefore, ensure relatively higher success rate of course recommendation and selection.

IV. APPLICATION OF PERSONALIZED RECOMMENDATION SYSTEM OF ONLINE COURSES

On the basis of the aforementioned design, this design will apply the personalized recommendation system to Shanghai Lifelong Learning Network to implement the recommendation system for online courses.

According to the functional design of the system, when a

learner registers information and completes courses on a learning platform, the database will generate user information table and learning behavior record, and form course recommendation according to the content of learning, learners' occupation or interest.

When a learner clicks on the learning resource of *Modern Life and Bank*, the system will recommend the course of *Introduction to Financial Knowledge* to him because this course is the rudimentary knowledge of *Modern Life and Bank*, and the learner might not learn it well or still have interest for further study. In addition, the occupation of the registered user is an engineer, and the system will recommend to the users *Engineering Mathematical Method and Introduction to Design Calculation* to guide the user to learn more. According to the information stored in the database, the interests of the users are diverse: science and technology, management and life, and the system recommends to the users the courses such as *Technology and Culture*, *Knowing Life Series and Promotion Decisions*, etc. It indicates that the personalized course recommendation system performs well in exploring and meeting the user demand. In this way, a recommendation based on the association rules of learners' learning data is formed.

In addition, in Shanghai Lifelong Learning Network, we can also, based on the combinational algorithm, recommend to the users the premium courses that are best rated by similar users. For example, when a user clicks to learn *Elderly Nutrition and Diet*, the course database will search the most similar users as well as the courses best rated in the collection of similar users by means of combinational algorithm, and recommend these courses to this user. Through algorithm-based search, we recommend to this user *Diet and Health in Four Seasons (Spring)*, *Life Source Series (14)*, *Poet Beggar*, *Tear-jerking*, etc. Through the course recommendation list, we can find that the all the recommended courses are about health care and preservation and well fitted with the user's interest. It also confirms that the personalized course recommendation system based on collaborative filtering algorithm has good recommendation effect.

The personalized resource recommendation system facilitates the users' course viewing, greatly saves the users' time to view and search the courses, and therefore, attracts more users to access the lifelong learning network. On the resource alliance platform of Shanghai Lifelong Learning Network, more than 80% trainees think that the learning network provides plentiful courses. Nearly 90% of the users recognize Shanghai Lifelong Learning Network, and they all express their plan to rely on the network for further study in the future.

V. CONCLUSION

With the development of society, economy, culture and public life, learner demand for premium and personalized education is increasing. Premium education requires good educational resources, and remote education plays an inestimable role in lifelong education. However, these premium resources are not necessarily what is urgently needed by each learner, and the learners need to reasonably use and screen

these resources according to their own needs. To find the required learning resources among the massive learning resources, it is essential to use digital learning resource recommendation, and resource recommendation provides the learners with personalized learning service in network-based environment, improves learning efficiency, satisfies the learning requirements of the learners, and offers learner-centered service in a real sense. The great development of Internet inevitably gives rise to information explosion and the application of personalized recommendation technology will surely meet the personalized learning requirements of learners to certain extent.

ACKNOWLEDGMENT

This paper is supported by "Shu Guang" award "MOOCs design and empirical research oriented Shanghai lifelong learning (13SG56)" and the Oriental Scholar program (TPKY052WMJ) from the Shanghai Municipal Education Commission and Shanghai Education Development Foundation. It is also supported by the 2014 Shanghai education scientific research key project "The Study of online learning mode for Shanghai lifelong learning (A1403)". Besides, thanks for the support of Science and Technology Commission of Shanghai Municipality research project "Shanghai Engineering Research Centre of Open Distance Education (13DZ2252200)".

REFERENCES

- [1] Yang Li-na. "Research for Promoting the Effect of Personalized Recommendation on Digital Learning Resources," *Modern Educational Technology*, vol.24, no.6, pp.84-91, 2014.
- [2] Sun Xin, Wang Guyong, Qiu Feiyue. "The research of personalized recommendation of online learning resources based on collaborative filtering recommendation technology," *Distance Education in China*, vol.8, pp. 78-82, 2012.
- [3] Chih-Ming Chen. "Intelligent web-based learning system with personalized learning path guidance," *Computers & Education*, vol. 51, pp. 787-814, 2008.
- [4] Wang Zhimei, Yang Fan. "Resource recommendation system based on similar learners exploitation," *Journal of Zhejiang University (Engineering Science)*, vol. 40, no.10, pp.1688-1691, Oct.2006.
- [5] Liu Zhiyong, Liu Lei, Liu Pingping, etc. "Learning resource personalizing recommendation based on semantic," *Journal of Jilin University (Engineering and Technology Edition)*, vol.39, pp.39-395, Sept. 2009.
- [6] Tom White, *Hadoop: The Definitive Guide*. Sebastopol: O'Reilly Media, Inc. June 2009.
- [7] Zhao Yanxia, Liang Changyong. "The Application of E-Commerce Recommendations Based on Association Rules," *Value Engineering*, no.5, pp. 88-91, 2006.
- [8] Zeng Chun, Xing Chunxiao, Zhou Lizhu. "A Personalized Search Algorithm by Using Content-Based Filtering," *Journal of Software*, vol. 14, no.5, pp. 999-1004, 2003.
- [9] Goldberg D, Nichols D, Oki. "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no.12, pp.61-70, 1992.
- [10] Apache. Hadoop Overview[EB/OL]. <http://hadoop.apache.org/common/docs/r0.20.203.0/>. 2011-04-05/2012-03-30.
- [11] Zhang Zhiguo, Liu Huaoliang, etc. "Research on video retrieval using high-level semantic," *Computer Engineering and Applications*, vol. 43, pp. 168-170, 2007.
- [12] Hu Shuangyan, Li Junshan, Li Jianjun. Video Retrieval Based on Latent Semantic Analysis. *Computer Engineering*, vol. 33, pp. 216-217, 2007.
- [13] Wengang C, De X. Content-based video retrieval using audio and visual clues. In *IEEE Proceeding of 2002 Region 10 Conference on Computers*,

- Communications, Control and Power Engineering*, Beijing, China, October 28-31, 2002, pp. 586-589.
- [14] Liu Jianguo, Zhou Tao, Wang Binghong. Research Progress of personalized recommendation system. *Progress in Natural Science*, vol.19, pp. 1-15, 2009.
- [15] Breese J, Hecherman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Medison, US (San Francisco: Morgan Kaufmann Publishers Inc.), pp. 43-52, July 24-26, 1998.
- [16] Deng ailin, Zhu Yangyong, Shi Bole. A collaborative filtering recommendation algorithm based on item rating prediction. *Journal of Software*, vol. 14, pp. 1621-1627, 2003.
- [17] Bong Resnick P, Iacovou N, SuchakM. Group lens: an open architecture for collaborative filtering of Netnews. *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, Chapel Hill, NC (New York: ACM), 175-186, October 22-26, 1994.

Jun Xiao is an associate professor of Shanghai Engineering Research Center of Open Distance Education, Shanghai Open University, China, the visiting scholar in department of computer science, San Diego State University, USA, and also the committee member of China Elearning Technology Standardization Committee. His research specialties focus on learning analytics, lifelong learning, digital learning system and educational resource repository research. He has led many large-scale Research and Development projects, such as Shanghai Educational Resource Center, Shanghai Lifelong Learning Network, Shanghai Learning Network, and this Cloud-Based Intelligent Learning System. He has published more than 30 articles on major publications, and he is the author of 5 books. Address for correspondence: Dr Jun Xiao, Learning Square, Room 505, No.288 Guoshun Rd, Shanghai, China. Tel: (+86) 021-25653263 Email: ecnuxj2003@163.com

Minjuan Wang is an oriental scholar at Shanghai International Studies University, China, a professor of Learning Design and Technology at San Diego State University, and a Program Manager for the Chancellor's office of California State University. Her research specialties focus on the sociocultural facets of online learning, and the design and development of mobile and intelligent learning. She has published peer-reviewed articles in *Educational Technology Research and Development*, *Computers and Education*, *Educational Media International*, *TechTrends*, and the *British Journal of Educational Technology*. She has also published book chapters on engaged learning in online problem solving, Cybergogy for interactive learning online, informal learning via the Internet, and effective learning in multicultural and multilingual classrooms. Address for correspondence: Dr MinjuanWang, 5500 Campanile Dr. PSFA 315, SDSU, San Diego, CA 92182-4561. Tel: 619-5943878 Email: mwang@mail.sdsu.edu.

Bingqian Jiang received the B.S. degrees in educational technology from East China Normal University. She is currently working towards the M.S. degree in the Department of Education Information Technology, East China Normal University. Her research interests include learning technologies, learning analytics, and lifelong learning. Address for correspondence: Ms. Bingqian Jiang, Room 405 Computer Building, ECNU, No. 3663 North East Zhongshan Rd., Shanghai, China. Tel: (+86) 021 25653454 Email: jbq6888@163.com

Junli Li is associate professor of Educational Technology, Shanghai International Studies University, and a core member of the Oriental Scholar team led by Minjuan Wang. She was also a visiting scholar of San Diego State University. Her research focuses on the design and development of mobile learning platforms. Address for correspondence: Dr Junli Li, Songjiang University Town, SISU, Shanghai, China. Email: ljshisu@163.com.

Theoretical Analysis and Experimental Evaluation of Bandwidth Amplification Attacks to Legitimate Websites

Dimitrios P. Iračleous, Kristofer E. Bourro, and Nikolaos Doukas

Abstract— Internet has turned into a vast field for increasingly high invasions and attackers, many of them using variations of DDoS (Distributed Denial of Service) attacks. In this work Domain Name Server (DNS) amplification attacks are considered as a variation of DDoS, are analysed and simulated in a testing environment. Comprehensive calculations are provided that given the attacker's exact network traffic to estimate to the size of the amplification factor and compare it to real measurements. Mitigation methodology is reviewed and further work is proposed.

Keywords— Amplification attacks, DNS flaws, theoretical analysis of attack.

I. INTRODUCTION

Distributed denial of service attack (DDoS attack) is an extended form of “denial of service attack” (DoS). In this work DDoS is produced by DNS amplified by a request initialized by one or multiple computers to specific targets [1]. These attacks are very difficult to prevent & to stop because they refer to responses on legitimate data coming from several servers. The simulation will include a list of 100.000 DNS resolvers. Afterwards a concerted and concurrent attack will be prepared using a line of 100Mbps for the beginning, including the requested metrics, and finally a line of 1 Gbps to compare with the previous repeated 3 times each. According to these metrics graphical representations are given and ways and methods to mitigate DNS amplification attacks are suggested.

II. LITERATURE REVIEW

DDoS attack is a malicious attempt of an attacker to use a host to attack to other hosts of gain access from them. In that case all affected hosts controlled from a single user which is the attacker's host. This can be achieved by taking advantage of security vulnerabilities or weaknesses and other disadvantages or misconfiguration settings of a system. After that (control gaining) the attacker will force the host to send huge amounts of data to a website or to another host and even

D.P. Iračleous is with Hellenic Military Academy, Vari, GR-16673, Greece (e-mail: dirakleous@ilabsse.gr).

K. E. Bourro is with IST College in collaboration with University of Hertfordshire, Pireos 72, 18346 Moschato, Greece.

N. Doukas is with Hellenic Military Academy, Vari, GR-16673, Greece (e-mail: ndoukas@ilabsse.gr).

send spam mails to particular email addresses. This kind of attack is called "distributed" because the attacker is using multiple hosts, including probably your own, to launch the denial-of-service attack [1-3].

A DDoS attack consists usually of two stages. At the beginning the attacker machine that has the role of the “intruder” looks after across the internet for vulnerable hosts or systems that can provide him access to install the attack tool and takes advantage of the most important resources of them for his own [2]. The DDoS master is a compromised system that can manage a number of other compromised systems with DDoS software like RAT, IRC or other HTTP programs. The next step as we can realize is for the handlers to find out a large number of infected computers to become the “compromised” computers that will also participate in the main attack in will increase rapidly the network traffic against the victim. The number of coordinated sources in a DDoS attack can vary from dozens to hundreds or even thousands. If all the above is achieved by the attacker then the whole structure of the attack is ready to fire against any targeted server. At this point it must be mentioned that DDoS attacks have a few more unique features apart from the volume of data that make effective defences extremely difficult and these are the ability of the attack packets to arrive from many different sources geographically distributed which makes IP source trace back also extremely difficult. Moreover an attack of that kind hasn't always have to be a strong or enormous attack because DDoS attack traffic will tend to appear legitimate and make filtering of the attack traffic a big headache without disrupting legitimate traffic on the victim server or website. Unfortunately today this kind of attack tools are available and can be easily find by anyone in the internet via official attackers Web pages or chat rooms of famous hack pages while finding a large number handlers and compromised computers has become technically trivial [3].

III. TYPES OF DDoS ATTACKS

Generally DDoS attacks are separated in three major types:

- Volume based attacks
- Protocol attacks
- Application Layer attacks

The first one includes UDP or ICMP floods, DNS amplification attacks and other spoofed-packet floods and they specialize in consuming the bandwidth of a victim. The

second includes SYN floods, fragmented packet attacks, Ping of Death, Smurf DDoS and more. This one is focused on real server’s resources and services through network protocols and services. The third one includes Slowloris, Zero-day DDoS attacks, DDoS attacks that target Apache, Windows or Open BSD vulnerabilities and more. The goal here is to crash the whole victim’s server.

Some common types include:

- User Datagram Protocol (UDP) flood attacks that leverages the session less networking protocol to flood random ports on a remote host with a lot of UDP packets.
- ICMP (Ping) flood which is almost similar to the UDP flood attack, an ICMP flood overwhelms the target resource with ping packets as fast as possible to consume both outgoing and incoming bandwidth.
- SYN flood attack exploits the known three-way handshake weakness in TCP connection sequence whenever a SYN request initiates a TCP connection.
- Ping of Death attacks make the attacker broadcast multiple malformed or malicious icmp (ping) packets to a host that exceeds the limit of 65,535 bytes including header per packet.
- Slowloris is a highly-targeted attack, enabling one web server to take down another host, without affecting other services or ports on the target network or server.
- Zero-day DDoS attacks are simply unknown or new attacks, where exploiting vulnerabilities for which no patch has yet been released [4].

There are certainly even more ways of DDoS attack types and methods but we try to stay in a minimal of the most famous reported.

IV. DNS AMPLIFICATION ATTACKS

The amplification DDOS technique consists of a small name lookup request to an open DNS with the source address spoofed to be another’s address. There are also many types of Amplification DDOS attacks like DNS, NTP, Chargen, XML-RPC Ping, and Smurf that attackers use to increase their attack size but just because the size of the response is typically considerably larger than the initial request, the attacker has easily the ability to amplify the amount of traffic directed at the other side (target). The common method of an attack is to overload the target system that eventually it cannot respond and finally become unavailable [5].

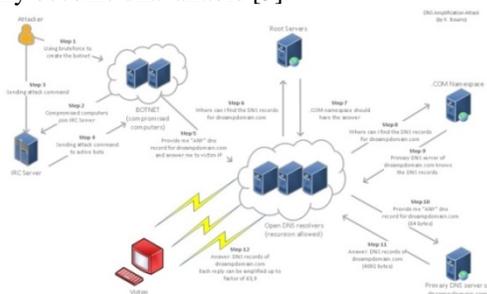


Fig. 1 DNS Amplification Attack steps

V. STEPS OF A DNS AMPLIFICATION ATTACK

The steps to implement DNS amplification attack

1. brute force to collect the compromised resources and create the army of botnets
2. makes the compromised computers joined the IRC server to be controlled remotely easier and execute all future command
3. sends the attack commands
4. attack commands send to all active bots.
5. active bots response to open resolvers to provide the “ANY” DNS record for dnsampdomain.com and spoof the answer to the victim’s ip.
6. open DNS resolvers (recursion allowed) requests the root servers to locate the list of DNS records for dnsampdomain.com.
7. root servers search and reply the .com name space servers for the possible answer. The root servers do not have the answer but they know where to redirect you in order to find what we are looking for.
8. open DNS resolvers requests the .com servers to locate the list of DNS records for dnsampdomain.com. The research continues with the next level of servers around the internet.
9. the .com servers replies the primary DNS servers of dnsampdomain.com who knows the answer (DNS records). It is the expected behaviour because they have the information needed and they provide the DNS resolvers.
10. the open DNS resolvers requests the primary DNS servers to locate the list of DNS records for dnsampdomain.com (64 bytes). The list will also be included in the attack script to the targeted site..
11. the primary DNS servers’ replies to the open DNS resolvers the DNS records of dnsampdomain.com (4092 bytes). The open DNS resolver has accepted the DNS request messages and they compose DNS response messages of the amplification record and return these to the systems that originated the request.
12. the open DNS resolvers’ redirects the answer of DNS records for the dnsampdomain.com to the victim’s site or host. At this final step the army starts to attack a targeted name server via the open recursive servers Each reply can be amplified up to amplification factor of 63,9 times. The targeted server tries to cope with the incoming load but it will eventually deny its service after a few while [6] [7].

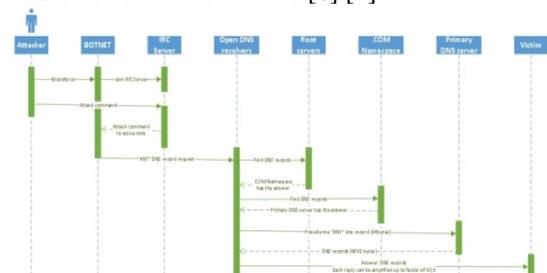


Fig. 2 Sequence of DNS Amplification Attack steps

VI. THEORETIC CALCULATION OF ATTACK TRAFFIC

At table 1 the required quantities are presented and described.

Symbols	Description	Units	Typical Value
B	Server Bandwidth	Bps	13.107.200
q	Size of request packet	Bytes/query	64
r	rate of queries	queries per sec	
a	Size of DNS response	Bytes per query	4092
T	Attack Traffic	Bps	
A	Amplification Factor		

Table 1 Algorithms symbols

Rate of queries is given by the formula

$$r = B/q. \tag{1}$$

Attack traffic is using the type:

$$T = r * a \tag{2}$$

So the amplification factor is given as

$$A = \frac{T}{B} \tag{3}$$

And then:

$$A = \frac{a}{q} \tag{4}$$

The amplification factor is independent of the server bandwidth. The theoretical maximum is about 64 [6].

Based on the theoretical approach the following calculations are straightforward.

As it can be seen the power of the network data is large and it is growing according the network bandwidth used each time. This is not always the case because there are losses in the lines and other problems that reduce these numbers. If we refer to the above metrics then we can see that the maximum of an amplification with no other traffic reducing is about x64 times larger but the average of an amplification is about x20 to x30 times which means that the 10 Gbps for example can produce output traffic to the targeted host at an amount of 200 Gbps in average and of course at a maximum of 639 Gbps which is highly difficult to achieve [9].

	100 Mbps	1 Gbps	10 Gbps
Rate of queries (queries/sec)	$r = \frac{13107200}{64} = 204800$	$r = \frac{134217728}{64} = 2097152$	$r = \frac{1342177280}{64} = 20971520$
Attack traffic (Bps)	$T = 204800 * 4092 = 838041600$	$T = 2097152 * 4092 = 8581545984$	$T = 20971520 * 4092 = 85815459840$
Amplification factor	$A = \frac{4092}{64} = 63,9375$	$A = \frac{4092}{64} = 63,9375$	$A = \frac{4092}{64} = 63,9375$

Table 2 Calculations based on the theoretical approach

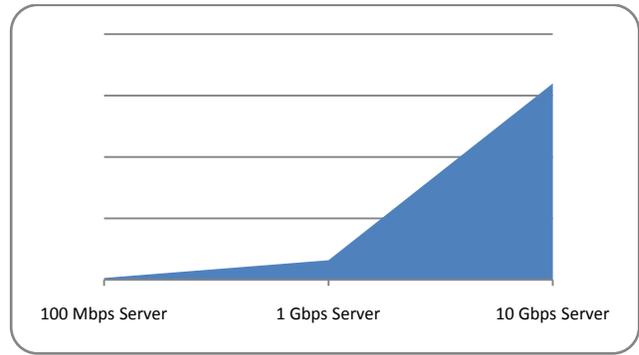


Fig. 3: The theoretical results of a full Amplification

VII. DESCRIPTION OF TESTING SYSTEM

A small network is set up consisting of two servers to be the part of the malicious hosts and they both run Linux Operating System the appropriate OS for such kind of actions in order to take the desired metrics for the measurements. The main issue was to find the servers that we could use to brute force to them and for that reason we choose to rent two servers to become our test servers to simulate the testing environment for our metrics. The first server hosts one dedicated 100Mbps network connection (bandwidth) and the second one a network line the order of 1Gbps. Both servers are having the option “IP spoofing” enabled. The target server used, who had to respond to the increased traffic was configured with DDOS protection enabled, for security and safety reasons, and all produced data came through the hardware firewall that was installed. Then we had to find a way to locate some open DNS resolvers to spread the attack and help collecting the required metrics afterwards. At the beginning we have run three attacks from the first server and then three attacks from the second server both targeting a third server that has hardware level DDOS protection so we were able to capture all the incoming traffic of those attacks from the hardware level.

A. Measurements

The measurements were accurate and here are the following results:

1) 100 Mbps server

The first set of three images relies on the first attempt of my amplification attack. When the first takes places you can notice how rapidly the amount of network traffic increases. The 100 mbps amplifies up to x36 times more when they hit the targeted host. This could very easily provide a full denial of a service delivered on a number of users [10].

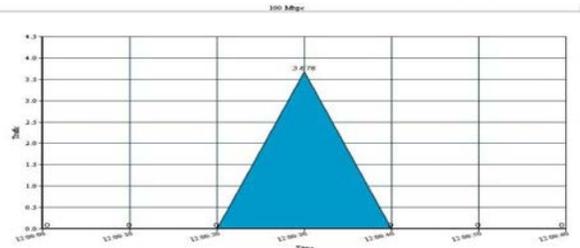


Fig. 4a 1st Measurement Attempt 100 Mbps

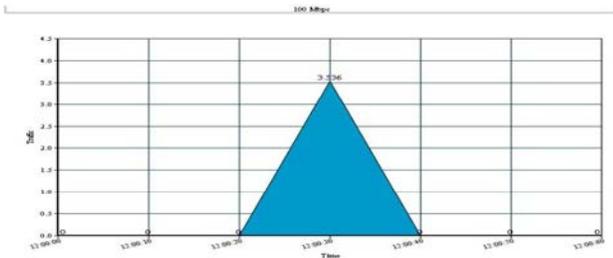


Fig. 4b 2nd Measurement Attempt 100 Mbps

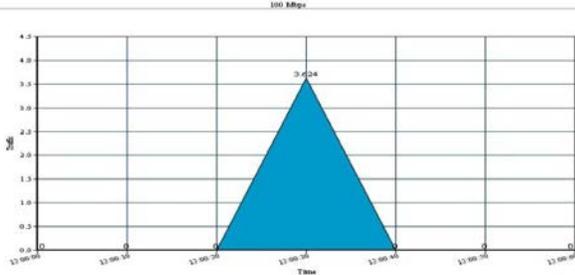


Fig. 4c 3rd Measurement Attempt 100 Mbps

2) 1 Gbps server

The second set of three images relies on the second attempt of the amplification attack. This attempt is a much bigger and more serious with much more harmful results to a targeted host.

If the bandwidth is increased from 100 mbps to 1 Gbps which is ten times in size the attack will not do the same with the result, for example from 36 times the attack will not reinforce to go on tenfold force such as 360 times stronger. This also shows the degree of difficulty and effectiveness.

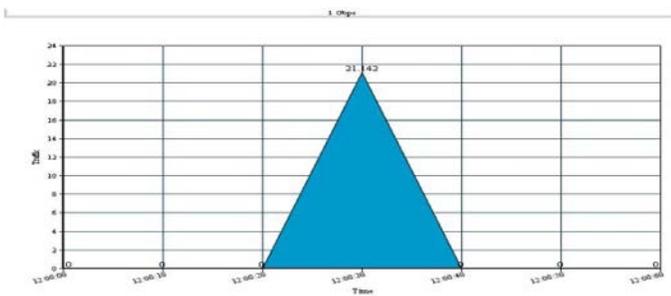


Fig. 5a 1st Measurement Attempt 1 Gbps

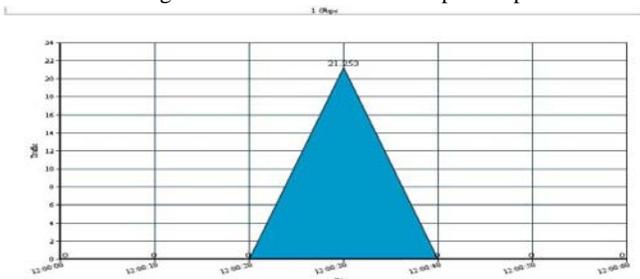


Fig. 5b 2nd Measurement Attempt 1 Gbps

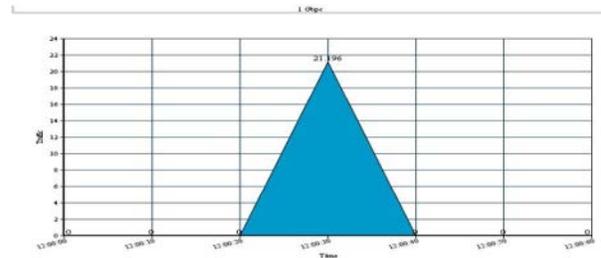


Fig. 5c 3rd Measurement Attempt 1 Gbps

B. Impact of the work

In the DNS attacks, the attacker uses an extension to the DNS protocol that enables large DNS messages, achieving the amplification and DNS data corruption effect. The attacker now issues a DNS request that he knows will evoke a very large response. Another characteristic in the end is the impersonation where each source host uses the targeted name server's IP address as its source IP address rather than its own.

The effect of spoofing IP addresses is responses to DNS requests which will be returned to the target rather than the spoofing hosts. Such kind of measurements and specifications shows why DNS amplification attacks differs from common DDOS attacks and why they are so extremely dangerous. They require little effort, creating vast overloading problems and performance degradation rapidly.

VIII. DETECTION AND PREDICTION

Besides it seems easy to detect an attack judging from the results it can produce, it is one of the most difficult issues anyone can solve. That happens because attackers use the reflection method and you can easily confuse authoritative from non-authoritative name servers in a DDoS amplification attack. One quick and reliable option we can use to detect or predict an attack is to refer on the internet and look after some free tools given to detect open recursive DNS resolvers. Such tools can scan entire network ranges and list the addresses of any identified open resolvers. These projects are:

- Open DNS Resolver Project (<http://openresolverproject.org>)
- The Measurement Factory (<http://dns.measurement-factory.com>)
- DNS Inspect (<http://www.dnsinspect.com>)

All the above projects give someone the ability to the user to search or check a name resolver in global scope and they also provide a variety of tools for administrators. The better way to predict an attack is being ready for an attack which means that DNS servers administrators should modify recursive name servers to accept queries from authoritative name servers and force source IP verification for not allowing spoofed address inside their network. Moreover we can use system tools in order to detect an attack if we are facing the symptoms of an attack or if we suspect such kind of behaviour.

IX. PROPOSED MITIGATION

For proper mitigation we should use a non-privileged source port (above 1024 port) for all DNS queries and to be much safer is better to use random and different each time source ports for every query. That means that authoritative name server operators should first deploy packet filters that drop traffic destined for the name server and having a source port equal to 53. We can only permit recursive queries to authorized users that belong in an organization and achieve limiting the recursion to authorized users reducing the risk for an attacker to gain this permission. For a network administrator the Response Rate Limit (RRL) by source IP address and the Response Rate Limit (RRL) by destination IP address are also two very special mitigating parameters. The first one gives the ability to set how many DNS queries per time can be accepted from an IP address or a subnet and the second one set the maximum number of queries that can reach any DNS server of an organization for example, dropping any additional network traffic above this threshold. Both mechanisms protect the DNS servers from overwhelming but in both cases it is recommended to have extensive tests before applying it, reducing unexpected problems in production and must have the proper infrastructure needed for such changes.

Another option that we have to consider first of all is our infrastructure and how it is designed from the beginning. If we are going to host the DNS server in our organization we must then configure a high availability solution to ensure our protection and the automatic failover scenario in case of an attack. So if an attack to our first server we can easily fail-over to the second one and decrease the impact to our business and users. Besides this we can also protect ourselves from attacks using techniques that configures or manages DNS state cache memory like pros and cons. Using long TTLs on our parent zone for certain delegation records is one of them which are associated with A or AAAA records. The value should set to a long one in order to exist in a recursive name server's cache for more time and reduce the impact of an attack because the content is cached and there is no reason to query it again and produce much more traffic. In case of a cloud-based DNS design for DDoS protection we can also redirect our DNS for our legitimate users to a new one and leave non-legitimate or attackers to our current DNS if we want to mitigate an attack of that kind. This solution requires cloud-based DDoS protection services and a series of action that should be deployed by the users after the redirection [10].

X. FURTHER WORK

DNS Amplification attacks are a fact today that is increasing day by day. DNS servers with recursive response option can be easily located by attackers across the internet and help them create huge amounts of network traffic while attacking. In this paper we try to analyse the role amplification factor, we try to calculate him, measure his affect and try to purpose protection and mitigation methods around it. As a future work we thought that we could use our testing environment and our metrics to create standard DNS attacking

stress test scenarios builder simulator for several websites and according to the results we will suggest mitigating measures or load protection configuration for the web site. Moreover we can create a serious DNS amplification measurement tool and provide it to customers who would like to test their web sites under certain circumstances on DDoS attacks. We can expand its capabilities with other types of DDoS and provide a full type report with solutions for the problems reported after the simulation of the attack. This report could also include all protection & mitigation methods suggested. These actions will not only take advantage of the whole project and my measurement but it will provide a productive solution for companies worldwide and their portals as well.

XI. CONCLUSION

In this work it has been demonstrated how big damage can be produced by a small amount of code and how exponentially this can be increased. Attackers find victims to flood their computers with a large number of network packets on a DDoS attack, taking advantage of security vulnerabilities or weaknesses for control gaining on multiple hosts and turn them into "compromised" before he launches the DDoS attack.

The exact produced network traffic of an attack depending on rate of the queries and the size of DNS response creating the amplification factor has been theoretically calculated and compared. Attackers prefer to use an extension to the DNS protocol that enables large DNS messages, achieving the amplification and DNS data corruption effect.

REFERENCES

- [1] M. Sheena, D. Madhuri and A. Annapurna, "Distributed Denial of Service Overview and Prevention," *Int. Journal of Computer Engineering & Applications*, vol. IV, no. 1/3, pp. 29-34, 2013.
- [2] US-CERT advisory, "<http://www.us-cert.gov/ncas/alerts/TA13-088A>," United States of America, 29 March 2013. [Online].
- [3] CloudFlare, "<http://blog.cloudflare.com/the-ddos-that-almost-broke-the-internet>," 27 March 2013. [Online].
- [4] C. Patrikakis, M. Masikos and O. Zourarakis, "Distributed Denial of Service Attacks," *The Internet Protocol Journal*, vol. 7, no. 4, pp. 13-35, 2004.
- [5] G. Kambourakis, T. Moschos, D. Geneiatakis and S. Gritzalis, "Detecting DNS Amplification Attacks," in *Critical Information Infrastructures Security*, J. Lopez and B. M. Hämmerli, Eds., Málaga, Spain, Springer Berlin Heidelberg, 2008, pp. 185-196.
- [6] G. Kambourakis, T. Moschos, D. Geneiatakis and S. Gritzalis, "A Fair Solution to DNS Amplification Attacks," in *Digital Forensics and Incident Analysis*, 2007. WDFIA 2007. Second International Workshop, Samos, 2007.
- [7] D.P. Iracleous, N. Papadakis, I. Rayies and P. Stavroulakis, "Cyber Warfare Scenario and Military Application," *2nd Int. Conf. on Applied and Computational Mathematics (ICACM '13)*, Athens, 2013, pp. 177-181.
- [8] E.V. Soroka, D.P. Iracleous, "Social Networks as a Platform for Distributed Dictionary Attack," *WSEAS Conference on Recent Researches in Communications and IT, Corfu*, 2011, pp. 101-105.
- [9] T. Deshpande, P. Katsaros, S. Basagiannis and S.A. Smolka, "Formal analysis of the DNS bandwidth amplification attack and its countermeasures using probabilistic model checking," In *High-Assurance Systems Eng.(HASE)*, vol. N2011 IEEE 13th Int. Symp. pp. 360-367, 2011.
- [10] C. Sun, B. Liu and L. Shi, "Efficient and low-cost hardware defense against DNS amplification attacks," In *Global Telecommunications Conference*, vol. IEEE GLOBECOM 2008, pp. 1-5, 2008.

Digital Image Segmentation Inspired by Carrier Immigration in Physical P-N Junction

Xiaodong Zhuang, Nikos E. Mastorakis

Abstract—In this paper, a new method for image segmentation is proposed, which is inspired by the carrier immigration mechanism in physical P-N junction. The carrier diffusing and drifting are simulated in the proposed model, which imitates the physical P-N junction. The effect of virtual carrier immigration in digital images is studied by experiments on test images and real world images. The sign distribution of net carrier at the model's balance state is exploited for region segmentation. The experimental results prove the effectiveness of the proposed method.

Keywords—Image segmentation, virtual carrier immigration, self balancing, physics inspired method

I. INTRODUCTION

IN physical P-N junction, the charge carriers in P-type and N-type semiconductor are holes and electrons respectively [1,2]. The density of hole is high in P-type semiconductor, while electron has a high density in N-type semiconductor. When the materials of the two are put together with compact contact, diffusion of carries will happen at the interface of contact due to their density difference (i.e. carrier moving from high-density side to low density side). Meanwhile, a space charge region is established due to the carrier diffusion and recombination. The space charge region will grow with more diffusing of carrier, but it in turn causes the drifting of carriers which is at the opposite direction of diffusing. A demonstration of the physical P-N junction is shown in Fig. 1. The above dynamic process will reach a balance state when the drifting and diffusing of carriers get balanced, and a stable P-N junction will be established under that dynamic equilibrium [1,2].

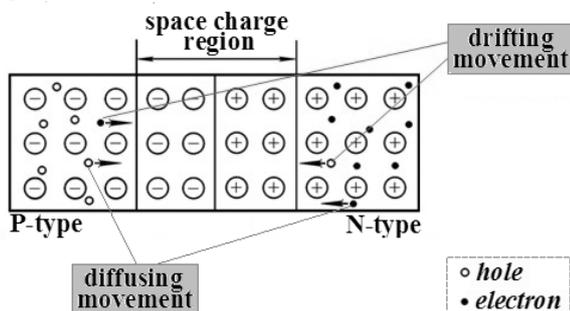


Fig. 1 The physical P-N junction

Image segmentation is a fundamental problem in image processing, which has significant value in both theoretic and practical research [3-5]. The basis of differentiation of two adjacent regions is their difference of image characteristics, such as grayscale, color, texture, etc. It is still an on-going and open research topic to segment image regions for various practical purposes [3-7]. In recent years, nature-inspired methods have attracted more and more research attention, in which the mechanisms in nature are imitated and adjusted in novel algorithms for image segmentation, and promising results have been obtained in such preliminary works [8-11]. In this paper, inspired by the physical P-N junction, the self-balancing mechanism of carrier diffusing and drifting is adopted in a novel segmentation framework, in which the region structure of image is formed by dynamic carrier diffusing and drifting.

In the following proposed model for region segmentation, the immigration of carrier caused by carrier density difference is called “diffusing”, and that caused by virtual electric field is called “drifting”. Of course, the formation of P-N junction is a microscopic physical process, while region segmentation is based on some algorithm implemented on computers. The algorithm proposed in this paper is inspired by the P-N junction, but not just a simulation of the physical process. In the following sections, the virtual carrier diffusing and drifting imitates the physical P-N junction, but the virtual electric field is defined artificially according to grayscale difference between adjacent image pixels. The proposed algorithm exploits the physical mechanism of carrier immigration, and introduces the factor of grayscale difference between the adjacent regions by the virtual electric field. Such difference between the algorithm and physical process should be noticed in reading the following sections.

II. THE MODEL OF VIRTUAL CARRIER IMMIGRATION IN DIGITAL IMAGES

The proposed model for image segmentation is as follows. Suppose there are two categories of virtual carriers: positive and negative, which imitates the physical electron and hole. On the 2D plane of digital images, each pixel is modeled as a container of virtual carriers. Each pixel has four adjacent pixels (except those on image borders), and correspondingly each carrier container has four adjacent containers. There is an interface between each pair of adjacent containers. Therefore, each pixel in the image corresponds to a carrier container with four different interfaces between its adjacent containers.

There are two features of the interface mentioned above. Firstly, the interface has permeability, which means the carriers at both sides of the surface can diffuse through it

Xiaodong Zhuang is with Qingdao University, Automation Engineering College, China (e-mail: xzhuang@worldses.org), and is also a cooperation researcher in Technical University of Sofia.

Nikos E. Mastorakis is with Technical University of Sofia, Industrial Engineering Department, Kliment Ohridski 8, Sofia, 1000 Bulgaria (email: mastor@wseas.org).

due to density difference. Secondly, there is a virtual electric field imposed on it, whose direction and intensity are determined by the grayscale difference between the corresponding two pixels connected by that interface. The virtual electric field is defined as:

$$e = K \cdot (g - g_a) \tag{1}$$

where e is the intensity of virtual electric field at the interface, K is a predefined positive coefficient, g is the grayscale of the pixel of interest and g_a is that of its adjacent pixel. The direction of the electric field is defined according to the grayscale relationship of the pixel pair: the side with higher grayscale value has the higher electric potential. In another word, the container corresponds to higher grayscale value will attract negative carriers, while that corresponds to lower grayscale attracts positive carriers.

Moreover, the effect of each virtual electric field is limited to its corresponding interface only, and does not influence other interfaces. The virtual electric field on each interface plays a key role in the formation of diffusing-drifting balance of virtual carriers in the model. In such a way, the model is established consisting of virtual containers of carrier (both positive and negative carrier), the interface between adjacent containers, the virtual electric field, and also the virtual carriers.

The evolution of the system is analyzed as follows. Initially, suppose the positive and negative carriers are of the same quantity, and each container has the same amount of carriers. Also suppose all the containers have the same volume, so that the density of carrier in a container is proportional to the amount of carrier in it. Therefore, there is no density difference of carriers between adjacent containers at that time. In another word, there is no carrier diffusion at the beginning. However, with the effect of the virtual electric field at each interface, the positive and negative carriers drift across the interfaces due to the virtual force applied by the electric field. The drifting then causes carrier density difference between two sides of the interface, which in turn makes the carriers to diffuse due to that density difference. Obviously, the diffusion has the opposite effect of drifting. For each interface and each container, such dynamic process evolves until a balance between drifting and diffusion is reached. The proposed model is shown in Fig. 2 and Fig. 3. Fig. 2 shows the details of carrier immigration between two adjacent containers, while Fig. 3 shows the overall structure of the model on digital image. The balance state is worth of study for possible use in image segmentation, and the net carrier (i.e. the resultant amount after the offset of positive and negative carriers) in each container is of value in further analysis.

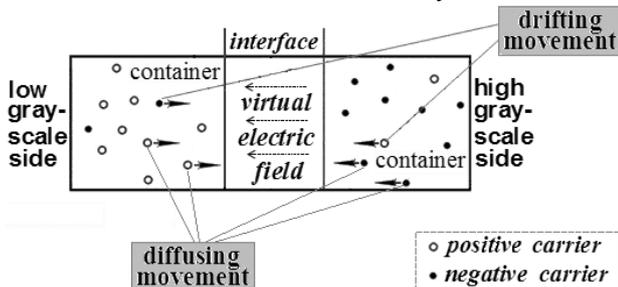


Fig. 2 Two adjacent containers in the proposed model of carrier immigration in digital images

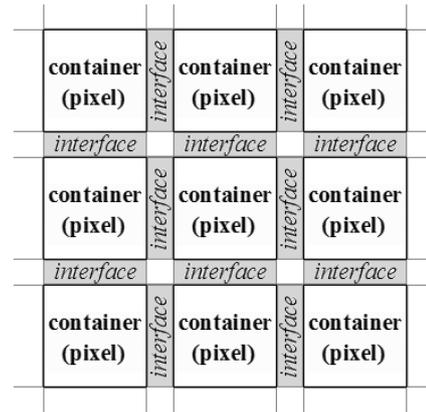


Fig. 3 The structure of the proposed model upon digital image

III. THE ANALYSIS OF VIRTUAL CARRIER IMMIGRATION

The effect of the above dynamic process can be analyzed for two typical cases. The first case is for the pair of adjacent pixels with large grayscale difference, such as pixels at different sides of the region border. According to Equation (1), because the intensity of the virtual electric field is defined as proportional to the grayscale difference, in this case the virtual electric field is strong. The strong field will cause the carriers to drift, which quickly causes obvious density difference of carriers between the two sides of the interface. And the density difference in turn causes the opposite movement - the diffusion of carrier. Therefore, in this case the drifting of carriers caused by the virtual electric field is the dominant factor of carrier movement, which increases the carrier density difference between the two adjacent containers at the region border. On the other hand, the diffusing of carrier is an induced one, which functions as a factor to balance the drifting. In a word, for the case of large grayscale difference between adjacent pixels, the difference of the carrier density in their corresponding containers has an increasing tendency. Since there are two kinds of carriers, and their movements under the virtual electric field are opposite, in the above case the net carrier (i.e. the net charge) in one container will become the opposite to the other. That is to say, in this case the net charge in one container will become positive, while that in the adjacent one will become negative.

The second case is for the pair of adjacent pixels with small grayscale difference, such as adjacent pixels within the same region. In this case, the virtual electric field at the interface is weak because of the small grayscale difference. Therefore, in this case the dominant factor is the diffusing of carrier due to carrier density difference. Because the dominant diffusing process will decrease the difference of carrier density, there is a tendency to an identical distribution of carrier within a region. As a result, the containers within the same region tend to have the same sign of net carrier (i.e. all positive or all negative).

The obvious difference between the above two cases just satisfies the requirement of segmenting adjacent regions. With the immigration of carriers going on, at the boundary of two different regions, the carrier density difference is large between the two sides of the region boundary. On the other hand, the carrier density of the containers within a region has a tendency to an identical distribution by the

diffusing process. The final result at balancing will reflect the differentiation of adjacent regions, which may provide useful clues for image segmentation. The simulation results in following sections provide experimental support to the above analysis.

There is difference between the proposed model and the physical P-N junction. In physics, the basic factor is the carrier density difference between the P-type and N-type semiconductors. The carrier density difference causes the diffusion of carrier, which forms the electric field (i.e. the space charge region) at the interface. The electric field in turn induces the drifting of the carrier. However, the process in the proposed model for region segmentation is just different. For each pair of adjacent pixel, the virtual electric field is firstly defined according to the grayscale difference between adjacent pixels, and keeps unchanged in the whole evolving process. The virtual electric field causes the drifting of carriers, which causes carrier density difference between adjacent pixels (virtual containers). Then the density difference in turn makes the carriers to diffuse, which is an opposite movement to drifting. Therefore, the proposed model adopts the drifting-diffusing mechanism of carrier in physics, but it is not just only a simulation of physical P-N junction.

IV. IMAGE SEGMENTATION BASED ON VIRTUAL CARRIER IMMIGRATION

A. Model Implementation by Computer Simulation

In the simulation of the model on computer, the simulation must be implemented in discrete steps (i.e. iteration by iteration). In one simulation step, the drifting speed of carrier (i.e. the amount of carrier immigrating from one container to the other in one iteration of simulation, or one simulation step) is defined directly proportional to the intensity of virtual electric field:

$$\Delta c_{drifting} = K_1' \cdot e \quad (2)$$

where $\Delta c_{drifting}$ is the amount of carrier drifting from one container into the other, K_1' is a predefined positive coefficient, e is the intensity of virtual electric field at the interface. According to Equation (1), $\Delta c_{drifting}$ is also proportional to the grayscale difference between the adjacent pixels:

$$\Delta c_{drifting} = K_1 \cdot (g - g_a) \quad (3)$$

where $\Delta c_{drifting}$ is the amount of carrier drifting from one container into the other, K_1 is a predefined positive coefficient, g is the grayscale of the pixel of interest and g_a is that of its adjacent pixel.

On the other hand, in one simulation step, the speed of diffusion has proportionality relationship with carrier density difference between the adjacent pixels. Suppose each container has the same size (or volume). Then the carrier density is proportional to the carrier amount in each container. Therefore, in the implementation the carrier density is substituted by carrier amount for diffusing:

$$\Delta c_{diffusing} = K_2 \cdot (c - c_a) \quad (4)$$

where $\Delta c_{diffusing}$ is the amount of carrier diffusing from one container into the other, K_2 is a predefined positive

coefficient, c is the net carrier amount in the container of interest and c_a is that in its adjacent container.

For Equation (4), there is another issue to be discussed in the implementation. There are two types of carriers in the model: positive and negative. Each container has both types in it. In the evolving process, the two types of carrier immigrate by drifting and diffusing respectively. For each container, the net carrier is the combination of the two types after the offset between them. For simplicity in implementation, the immigration of carriers is measured by the amount of net carrier. In another word, the carrier density and the flow of carrier between containers are measured by net carrier amount.

The steps of implementing the model are as follows. At the beginning, the amounts of positive and negative carriers are equal in each container. Also suppose the amount is sufficient for arbitrary amount of carrier immigration in the simulation. Then the virtual electric field is calculated at each interface between adjacent containers, which is proportional to the grayscale difference between corresponding adjacent pixels. The detailed simulation step is as follows.

Step1 For each of the four interfaces of every virtual container (or pixel), do the following: calculate the drifting amount of carrier due to virtual electric field; calculate the diffusing amount of carrier due to carrier density difference; sum the above two for all the 4 interfaces of a container to get its total change of net carrier amount; update the net carrier amount in that container;

Step2 After all the containers update their net carrier amount, calculate the average change of net carrier for all the containers. If the average change of net carrier is smaller than a predefined threshold, it is close enough to the balance state, and the simulation stops; otherwise, return to **Step1** to begin a new iteration of simulation.

B. Simulation Results for Simple Test Images

A typical one of the experimental results for a series of test images are shown in Fig.4. The simulation is implemented by programming in C. Fig. 4(a) shows a test image with a rectangle region. In order to demonstrate the process of carrier immigration step by step, the intermediate results at several specific simulation step numbers are recorded together with the final result. The intermediate results are shown from Fig. 4(b) to Fig. 4(f). In the following results of net carrier sign distribution such as Fig. 4(b) to Fig. 4(f), the white and black points represent the positive and negative sign of net carrier respectively, and the gray points represent that the net carrier is zero (i.e. not a definite sign yet). In Fig. 4(b), it can be seen that shortly after the simulation starts, only the points close to the region border have a definite sign of net carrier, From Fig. 4(b) to Fig. 4(f), it is clear that the positive sign part expands inward, while the negative sign part expands outward from the rectangle border with increasing simulation time. The final result of sign distribution of net carrier in Fig. 4(g) can provide a definite segmentation of Test image 1.

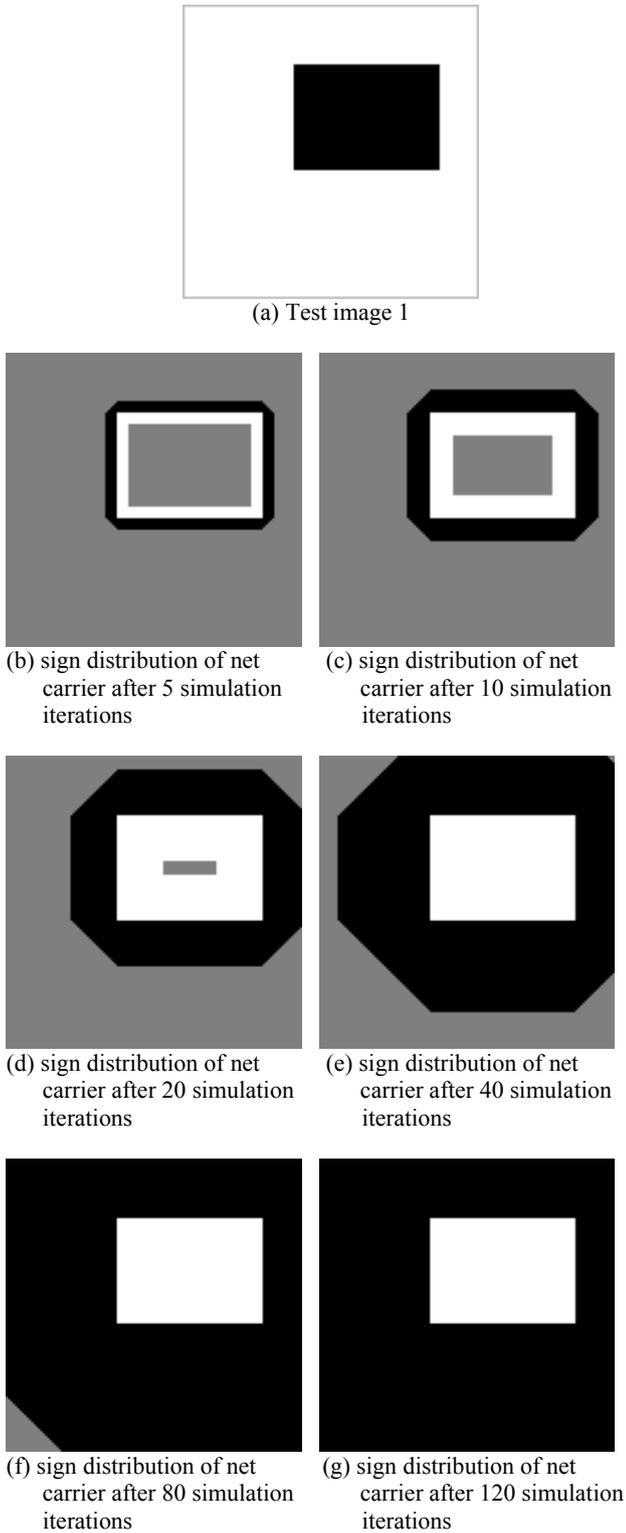


Fig. 4 The simulation results for Test image 1

The above interesting results for test images can be analyzed as follows³. At region borders with large grayscale variation, the dominant factor of carrier immigration is the drifting caused by the strong virtual electric field. As the overall effect, for a pair of containers, the positive carriers tend to gather into one container, while the negative carriers gather into the other. If the electric field is defined as proportional to the grayscale difference, the container corresponding to higher grayscale pixel has the higher electric potential, which attracts negative carriers

to gather. On the other hand, the positive carriers are attracted into the other container corresponding to lower grayscale. Such effect will increase the difference of net carrier amount between the adjacent containers, and the sign of the net carriers in them also tend to become opposite.

On the other hand, for a local area within a region, the grayscale difference is small. Correspondingly, the virtual electric field is also small. Here the dominant factor for carrier immigration is the density difference of carriers between adjacent containers, which makes the carriers diffusing, and produces a tendency of identical carrier distribution inside a region. As the result, the amount and sign of net carrier for the containers inside a region tend to become homogeneous. Therefore, the global distribution of the sign of net carriers at balance state can provide effective basis for image segmentation.

C. Image Segmentation Based on the Proposed Model for Real World Images

In the above experimental results for the test images, it is shown that the sign of net carrier are opposite in two adjacent regions, which can provide the basis of region division in images. In order to obtain the segmentation result from the sign distribution of the net carrier, a region grouping approach is proposed as following:

Step1: Implement the simulation of the virtual carrier immigration as proposed in section 4.1;

Step2: Obtain the sign distribution of the net carrier;

Step3: Group the adjacent containers (i.e. image points) with the same sign of net carrier as connected points in same region. In the region grouping process, the adjacent pixels of the 4-connection (i.e. the upper, lower, left and right pixels) for an image point p is investigated. If any of the four adjacent containers (pixels) has the same sign of net carrier as p , it is grouped into the region which p belongs to. The obtained regions are the result of region segmentation for the image.

The obtained set of regions is the result of region segmentation. Fig.5 shows the region segmentation results according to Fig. 4(g), where different regions are represented by different gray-scale values.



Fig. 5 The region grouping result for Fig. 4(g)

However, real world images consist of more complex regions than the simple test images. Moreover, gradual grayscale changes are commonly seen in natural images rather than sharp grayscale change at the region borders. To investigate the effect of the carrier immigration method on real world images, experiments are carried out for a series of real world images. For demonstration, some of the results

are shown in Fig. 6 to Fig. 9, which are for the house image, the peppers image and the medical heart image. The experimental results indicate that the proposed method can obtain large amount of regions (more than a hundred) because of the complexity of real world images. There are 537 regions obtained for the house image, 177 for the peppers image, and 533 for the medical heart image. The regions in the results are shown in Fig. 6(c), Fig. 7(c), Fig. 8(c) respectively.

To obtain practically useful segmentation result, a region merging method is proposed for the above region segmentation results based on the gray-scale similarity of adjacent regions. First, an expected number of remaining regions after merging is given (usually by trail). Then the following steps are carried out to merge regions until the expected region number is reached:

- Step1:** For each region in the image, calculate its average gray-scale value.
- Step2:** Find the pair of neighboring regions with the least difference of the average gray-scale, and merge them into one region.
- Step3:** If current region number is larger than the expected region number, return to **Step1**; otherwise, end the merging process.

For each real world image, the figures show the original image, the sign distribution of net carrier at the balance state, the region segmentation results by grouping, and also the result of region merging. In the sign distribution of net carrier, the white points represent positive net carrier, and black points represent negative net carrier. In the region segmentation results and region merging results, different regions are represented by different grayscale values.

For the house image, the remained region number after merging is 50 in Fig. 6(d).

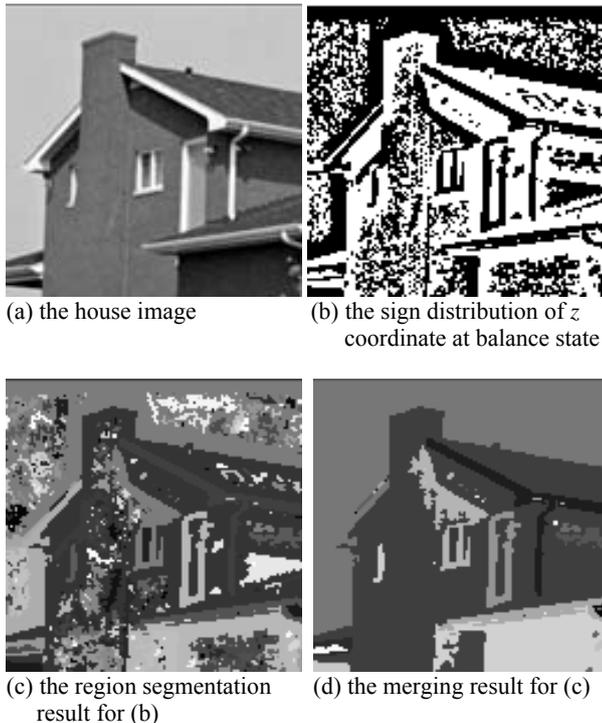


Fig. 6 The experimental results for the house image

For the peppers image, the remained region number after merging is 50 in Fig. 7(d).

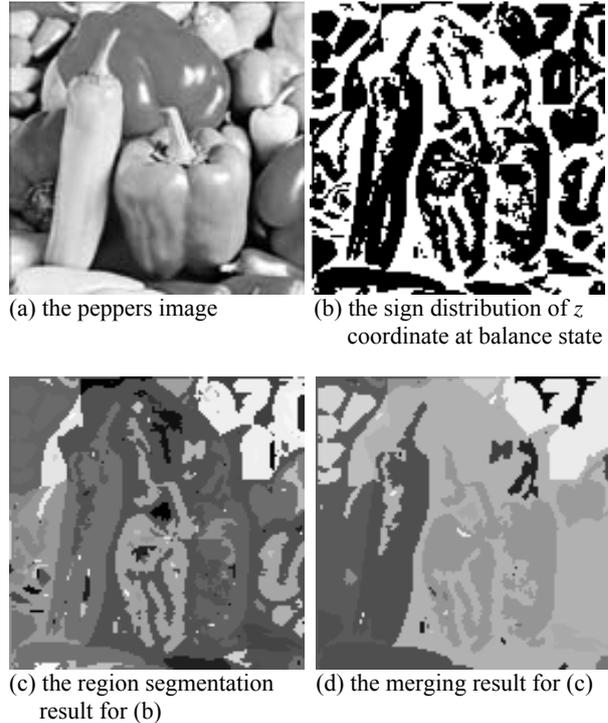
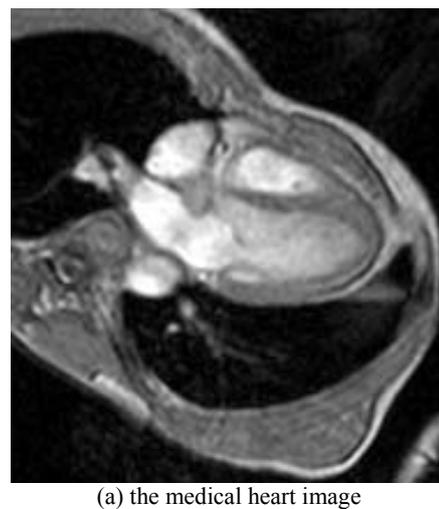


Fig. 7 The experimental results for the peppers image

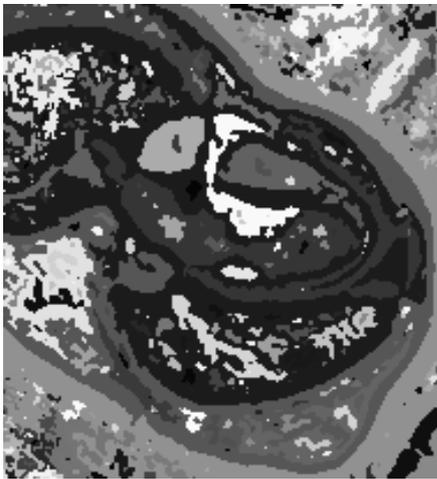
For the medical image of the heart, the remained region number after merging is 50 in Fig. 8(d). Fig. 8(d) shows the heart structure clearly. Moreover, the average of the net carrier change for all the points is calculated and recorded as a measurement of the convergence degree to the balance state. Fig. 9 shows the relationship between that average value and the simulation time, which indicates that the process of carrier immigration approaches the balance state with the simulation going on.



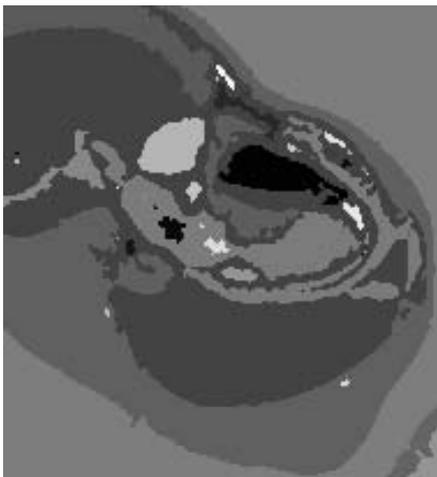
(a) the medical heart image



(b) the sign distribution of z coordinate at balance state



(c) the region segmentation for (b)



(d) the merging result for (c)

Fig. 8 The experimental results for the medical heart image

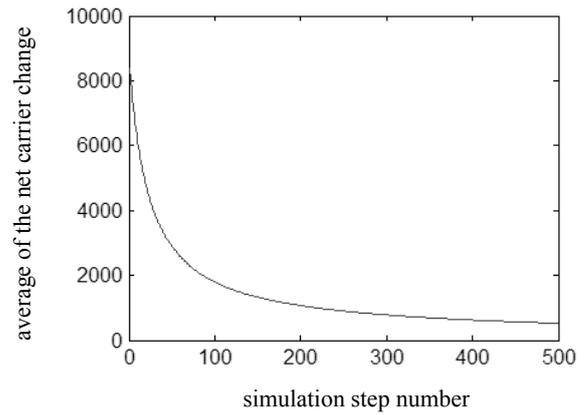


Fig. 9 The relationship between the average change of net carrier and the simulation time (for the medical heart image)

The above experimental results prove that the analysis of the model's dynamics in section 3 still holds for real world images. The proposed method is effective in segmentation of real world images. Relatively large amount of regions can be obtained by the grouping of the net carrier sign due to the complexity of real world images. From the results, it can be seen that main object regions can be well segmented, and some regions are segmented in perfect detail. However, at some part of object boundaries, two objects are not well separated due to reasons like grayscale similarity. It is indicated that other image features besides grayscale may be introduced into segmentation for improvements.

V. CONCLUSION

A new model of virtual carrier immigration on digital image is presented by imitating the diffusing and drifting of carriers in physical P-N junction. The virtual electric field between adjacent pixels is defined according to their grayscale difference, which is the major difference between the proposed model and real P-N junction. In the model, the two kinds of immigration are: carrier drifting caused by the virtual electric field, and the carrier diffusing caused by the carrier density difference. The direct local interaction and indirect global interaction of the above two carrier movements can lead to a balance state of carrier distribution, which provides clues for region segmentation. Image segmentation is implemented based on the sign distribution of net carrier at balance state, and a merging step is applied to get more comprehensible and useful segmentation results. The experimental results for test images and real world images prove the effectiveness of the proposed method.

For improvement of the segmentation results, experiments will be carried out to investigate the segmentation results under various parameter configurations in the method. Color and texture feature will also be introduced into segmentation for possible improvement. And more reasonable merging process will also be explored as a necessary post-processing step to obtain better results of region merging.

REFERENCES

- [1] Steven H. Simon, *The Oxford Solid State Basics*, Oxford University Press, August 16, 2013.
- [2] Walter A. Harrison, *Solid State Theory*, Dover Publications, January 20, 2011.
- [3] Yujin Zhang, *Advances in Image And Video Segmentation*, IRM Press, May 2, 2006.
- [4] I.V. Gribkov, P.P. Koltsov, N.V. Kotovich, A.A. Kravchenko, A.S. Koutshev, A.S. Osipov, A.V. Zakharov, Testing of image segmentation methods, *WSEAS Transactions on Signal Processing*, Vol. 4, No. 8, pp. 494-503, 2008.
- [5] N. Senthilkumaran, R. Rajesh, Image Segmentation – A Survey of Soft Computing Approaches, *Advances in Recent Technologies in Communication and Computing*, (ARTCom '09), pp. 844-846, 2009.
- [6] A. Gacsádi, L. Tepelea, I. Gavrilut, O. Straciuc, Energy based medical imaging segmentation methods by using Cellular Neural Networks, *Proceedings of the 15th WSEAS International Conference on Systems*, pp. 190-195, 2011.
- [7] Intan Aidha Yusoff, Nor Ashidi Mat Isa, Two-dimensional clustering algorithms for image segmentation, *WSEAS Transactions on Computers*, Vol. 10, No. 10, pp. 332-342, 2011.
- [8] Mark S. Nixon, Xin U. Liu, Cem Direkoglu, David J. Hurley, On using physical analogies for feature and shape extraction in computer vision, *Computer Journal*, Vol. 54, No. 1, pp. 11-25, 2011.
- [9] X. D. Zhuang, N. E. Mastorakis, Image analysis based on the discrete magnetic field generated by the virtual edge current in digital images, *WSEAS Transactions on Computers*, Vol. 10, No. 8, pp. 259-273, 2011.
- [10] X. D. Zhuang, N. E. Mastorakis, Image sequence analysis based on the 3D relative potential inspired by Physical Electro-Static Field, *WSEAS Transactions on Computers*, Vol. 11, No. 10, pp. 349-365, 2012.
- [11] X. D. Zhuang, Nikos E. Mastorakis, The relative potential field as a novel physics-inspired method for image analysis, *WSEAS Transactions on Computers*, Vol. 9, No. 10, pp. 1086-1097, 2010.

Designing Engaging Mobile Learning for K-12 Classrooms

Minjuan Wang, Melissa Calderwood, Yong Chen, Junli Li

Abstract— This paper addresses mobile computing with a focus on mobile learning. Applying a motivational model (ARCS), the research team conducted a survey to examine the current use of mobile learning to support the millennial learners in the K-12 classroom. Based on the survey findings, researchers explore how to design mobile learning lessons for schools by motivating the young learners. At the end, this report provides recommendations on how to improve the instructional design for mobile learning.

Keywords— mobile learning, mobile computing, instructional design, millennial learners

I. INTRODUCTION

SOcial connectivity and the emergence of wireless, broadband internet through affordable devices challenge educators to rethink the possible. From a technological perspective, society has “changed out of all recognition in the past few decades” [1]. Access to the global marketplace, social meeting places and multimedia has become increasingly available via mobile devices. As a result, educators, administrators and instructional designers are challenged to stretch the boundaries of the classroom and introduce content to learners in a context that makes it immediately relevant, usable and of late, mobile for them.

II. LITERATURE REVIEW

The explosion of mobile communications technology and the resultant portable lifestyle introduced into society generated the latest learning style, mobile learning (mLearning), which challenges the traditional model of classroom education. Current literature suggests the textbook and lecture instructional model, featuring a teacher as the central disseminator of learning, is quickly becoming obsolete in lieu of a model where students access, use and create content in real world contexts [2].

Schubert highlights interesting data that justifies and helps in forecasting the growth of an increasingly mobile student and classroom [2]. His research states that today’s teenager “sends more than 3,000 text messages each month.” He also claims that there are “approximately 500 television and 10,500 radio stations, about 350,000 iPhone applications and more than one trillion web pages,” indicating the evolution of information

availability [2]. The integration of this information affords educators an opportunity to immerse students in rich content with a flexible access point they’re comfortable with and enthusiastic about. This study will summarize the literature researched and further explore the implications of mLearning within three narrowly focused themes: (1) the definition of mobile learning; (2) current use of mLearning to support the millennial learners in the K-12 classroom; (3) learner motivation and mLearning lesson design for the K-12 classroom.

A. Mobile Learning (mLearning)

The most widely accepted definitions of mLearning center on the general theme of learning through the use of mobile devices. However, many sub-themes also exist and indicate that mLearning is focused less on the device and more on its use during the learning activity: (1) learning involves the use of a mobile web enabled multimedia device that allows for access, use and creation of content anywhere, at any time [3]; (2) the learning activity is flexible and encourages learners to draw on various resources including social collaboration with peers, accomplished performers, coaches and teachers [4]; (3) the technology, the learner and the learning are mobile and make content accessible in a usable context when it is immediately relevant [5].

B. Impacts of Mobile Learning on the Millennial Learners

McMahon and Pospisil [6] describe millennial learners as a group focused on social interaction, highly accomplished at multi-tasking and adept with the use of technology in social media, learning and workplace applications. They also specify that the millennial prefers group activities, embraces information technology and requires rapid access to information [6]. Given these preferences and the continued infusion of technology in all aspects of modern culture, significant attention is given to the adequacy of learning activities in supporting the preferences of the information literate generation in schools and entering the workforce. Several overlaps exist between the design of Mobile Learning and the preferences of the millennial learner including: (1) learner centric activity that incorporates rich experiences relevant to the learner’s job or personal interests and goals [4]; (2) immediacy of results and individualization of feedback [7]; (3) flexible, on-demand design which enables freedom of choice and removes the tether to the instructor [1]; (4) collaboration and social interaction [1].

C. Learning Motivation and Mobile Learning

Keller's theory [8] for learner motivation and model for motivational design was used as a framework for the motivational aspect of this study. Keller's model [8] provides a systematic approach to "designing the motivational aspects of learning environments to stimulate and sustain students' motivation to learn." The four major components of Keller's ARCS model [8] include: Attention, Relevance, Confidence and Satisfaction. The Attention step focuses on capturing the learner's attention and maintaining it throughout the course of instruction. The second step, relevance, notes that once the learner's attention is gained, it is important that the learner then identify how the material relates to his or her interests and life. The confidence category describes that a learner will feel motivated if he or she is able to demonstrate some measure of competence related to the material being presented. Lastly, in the Satisfaction step Keller states the learner will be pleased if he or she feels value in what he or she has accomplished as result of the instruction.

Recent studies examine motivation in technology based instructional environments. The first study examined was conducted by scholars Ju-Ling Shih, Hui-Chun Chu, Gwo-Jen Hwang, and Kinshuk [9]. The researchers examined the impact of PDAs in an elementary science classroom in Singapore. They also examined student and teacher attitudes towards context-aware ubiquitous learning. In this study, fifth-grade students used PDA's and a wireless network to identify vegetation on the school's campus. The self-paced lesson challenged students to explore the campus, identify vegetation and construct knowledge by accessing relevant content via the PDA. The research team found increased motivation in four areas: (1) increased personalization; (2) active participation; (3) interaction with learning content in context; (4) construction of knowledge without the direct influence of the teacher.

In a second study, Warschauer, Cotton, and Ames [10] studied how laptops and a web blog helped middle school students in Colorado become more inspired writers. Students completed all writing assignments on an issued laptop; in addition, they regularly contributed to a blog where they shared their digital writing experiences. Upon analysis of the blog contributions, researchers noted the following themes regarding the students' comparison of the laptop writing to handwritten work: (1) ease of access to editing tools; (2) improved accessibility of information; (3) increased sharing and collaboration; (4) increased motivation as a result of increased access and visibility on their work.

Each study indicates that learner motivation is impacted by the integration of m-learning technology and lesson design and report findings apparently aligned with the core principles of m-learning as well as the preferences of the millennial learner. However, specific research examining the infusion of Keller's motivational design model with lesson development for m-Learning in the K-12 classroom is presently limited and will be further explored in this study.

D. Contextual Factors

Several limitations hindered the researchers' ability to explore mLearning in K-12 classroom settings to the depth originally intended.

- 1) *Teacher Participation.* All participants completed the online survey, but they did not volunteer to share their mobile learning lesson plans for the content analysis component of the case study. Because of this, researchers resorted to a Google search to complete this task. While analyzing mLearning lesson plans from different states and school sites, the researchers observed that there is not one uniform template for lesson design. Some school districts require explicit reference to state academic standards, while others do not. This complication will be explored further in the findings section of the report.
- 2) *Duration of Study.* If this were an ongoing study, developed over several academic semesters, the researchers would benefit from observing and interviewing students who are actually using mobile devices in their classrooms.
- 3) *Geographic Factors.* The researchers were unable to access school sites within their geographic areas in which they could visit to collect data for this case study. The researchers felt that actually observing and having face to face conversations with mLearning teachers, instead of surveying them long distance, would provide more insight on their methodologies and daily implementation of successful mLearning classrooms.

III. METHODOLOGY

This section presents a detailed explanation of the evaluation framework used, participants targeted, data collection instruments, and data analysis procedures.

A. Evaluation Framework

Current population trends and shifts in student expectations and performance in learning environments suggest that the emergence of the millennial Generation in classrooms. This has necessitated an evaluation of how educational design is meeting the needs of a new generation of learners who are coming of age in an era marked by societal connectivity. In an effort to quantify the impact of mobile learning on learner motivation, this study focuses on the incorporation of learner motivation principles in curriculum development and an analysis of lesson plans used to guide learners in mLearning exercises. This study will inform stakeholders and fuel the design of appropriate and functional lesson plans by enhancing situational awareness, validating the effectiveness of the ARCS Model, and identifying measures for effectively incorporating mLearning into the traditional classroom. The approach must be a blend of both the formative and summative aspects of the clarificative evaluation [11]. In this case, current trends in K-12 education suggest that learner motivation will be cultivated through the introduction of mLearning devices rather than the content of the lesson plans developed to support educational objectives.

Lesson plan review, participant input, and literature suggest that an accurate measure of learner motivation in mLearning exercises may not exist in a formal context. This clarificative evaluation will examine: (1) alignment of mLearning educational objectives with state mandated K-12 learning standards; (2) the consideration of Keller's motivational design principles in the development of mLearning lesson plans; (3) the impact of educator attitudes and technological support in the development of mLearning exercises.

B. Data Collection Instruments

Data collection for this evaluation was conducted in a two-fold manner by which the research team solicited and analyzed K-12 educators involved in mLearning programs and through the analysis of lesson plans by the research team utilizing a content coding rubric divided into the elements of the ARCS model.

C. Educator Survey

The survey developed for this evaluation was designed to focus heavily on teacher and student expectations and their level of preparation for the use of mobile technology in the classroom and the educator's incorporation of Keller's ARCS model in the development of mLearning lesson plans. The survey is provided in full in the appendices and the two focal points of the survey are discussed further below.

- 1) *Preparation and Expectations for use of mobile technology in the classroom:* This section provides data on the level of preparation and support that educators receive while incorporating mobile technology into their classrooms. This information may yield critical data points needed to enable the research team to further isolate and target the shortcomings associated with the design and implementation of mLearning educational objectives. The participants will provide a baseline for understanding the support mechanisms currently in place or needed to ensure a transition to mobile learning technologies that is well received by students and educators alike.
- 2) *Keller's ARCS Model for mLearning lesson plan development:* The second focal point of the survey was designed to gauge whether educators recognize and incorporate the principles of Keller's ARCS model for learner motivation into mLearning exercises. This information, when correlated with the first component of the survey, may yield data that can support the development of learner centric mLearning activities that promote learner motivation and support the educational objectives of the specific K-12 element. The research team expects that there will be a correlation between the level of comfort with technology and the consideration of ARCS principles in the design and implementation of mLearning activities.

D. Lesson Plan Analysis

The research team conducted a concurrent analysis of a sample of lesson plans designed for mLearning programs across all levels of the K-12 curriculum, utilizing a content coding

rubric. The elements of the sampled lesson plans were scored from (1) Highly Ineffective through (5) – Highly Effective. The content coding rubric is provided in full in the appendices and the major components of the coding rubric are discussed in detail below.

- 1) *Content:* In this section of the coding rubric the research team quantified the degree to which the lesson plan incorporates components of effective development including: standards, objectives, and scaffolding. In particular, the research team is looking to correlate data on achieving state mandated curriculum objectives through the use of mLearning exercises.
- 2) *Motivation:* This component of the coding rubric incorporates the dimensions of Keller's ARCS model. Each of the lesson plans will be analyzed to determine how effectively the mLearning initiative incorporates Attention, Relevance, Confidence, and Satisfaction. In particular, the research team will investigate how the use of instructional media is used to satisfy the elements of learner motivation.
- 3) *Assessment:* This section of the coding rubric affords the research team the opportunity to correlate the data from the Content and Motivation sections with defining specific assessment credentials. This section should validate whether mLearning is an effective means to assess students against mandated state curriculum standards.
- 4) *Technical:* The final component of the coding rubric allows the research team to delineate the objectives of the mLearning lesson plan. The incorporation of mobile technology in curriculum development should not shift the focus of learning objectives. When analyzing the lesson plans, the research team is seeking to validate that the use of technology is not clouding the desired learning outcome.

E. Participants

Survey participation and lesson plan analysis were largely dependent on access. Because the survey is administered electronically, delivery requires the record of the respondent's e-mail address. In an effort to ensure adequate distribution to a population currently involved with mLearning applications, the research team leveraged the professional network of one of the team members in three states. The research team did investigate the possibility of disseminating the survey to various school districts in an effort to enhance the survey population. The team ruled out this approach in an effort to ensure that the survey was limited to educators incorporating mobile learning technology into their curriculum. Limiting survey responses to only those with mLearning experiences limits the diversity of the respondent pool but preserves the integrity of the response.

With the survey participants identified, the research team shifted the focus to gaining access to lesson plans to assess against the content coding rubric. The availability of lesson plans was a limiting factor with the analysis. The majority of educators identified for the survey either had personal or institutional restrictions that precluded them from submitting mLearning lesson plans for analysis. Though incongruent to

the survey respondent pool, the research team analyzed public domain lesson plans that incorporated various instructional media.

IV. FINDINGS AND ANALYSIS

The findings and analysis segment of this report is divided into two sections focused around different aspects of learner motivation in mobile learning. The first section describes the observations, intentions and attitudes of educators currently supporting their curriculum with the incorporation of mobile wireless technology in the classroom. This section will help in evaluating the preparedness of instructors as well as students and schools for mobile learning. The second portion of the findings include a content analysis of a sample of lesson plans used to guide learner centered mobile learning activities in the K-12 classroom. This section will be used to evaluate the integration of Keller’s motivational design principles into a set of mobile learning activities.

A. Observations and Attitudes of Educators

Data in this section was collected from a sample of nine educators in the United States including representatives from California, Missouri, and Washington. Although the sample size is small and does not represent the high school grade levels, the population includes teachers from a variety of grades and incorporates representation from all core subject areas (math, science, social studies, language arts, and English language learners) with the exception of physical education. Participants in the survey ranged in experience from first term teachers through tenured faculty members with more than fifteen years in the classroom; the mean teaching experience for the sample group was 4-6 years. The survey respondents are classified in Table 1.

Table 1. Teacher experience

Respondent Identification Number	State Currently Teaching	Years of Teaching Experience	Subject(s) mLearning Applied	Current Grade Level
R1	WA	1-3 years	Science	grades 6-8
R2	CA	7-10 years	Language Arts	grades 6-8
R3	CA	4-6 years	Social Studies	grades 6-8
R4	MO	15+ years	All Areas	grades 6-8
R5	CA	4-6 years	Math	grades 6-8
R6	WA	15+ years	Math	grades 1-2
R7	CA	7-10 years	Math	grades 6-8
R8	WA	4-6 years	All Subjects	grades 1-2
R9	WA	4-6 years	no response	grades 1-2
R10	CA	4-6 years	Language	grades 6-8

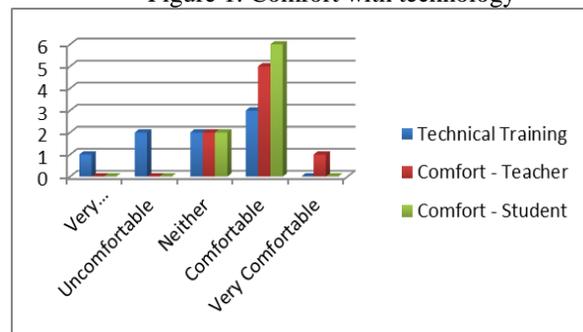
			Arts	
--	--	--	------	--

The study instrument was used to indicate teacher and student expectations and their level of preparation for the use of mobile technology in the classroom. The data specifically represents respondent’s answers to a variety of questions beginning with three subcategories of the prompt:

Please rate how prepared you and your students are for the use of mobile wireless technology in the classroom: (1) my level of training in developing coursework using mobile technology; (2) my comfort level with using mobile wireless technology in the classroom; (3) my student’s comfort level with using mobile wireless technology in the classroom.

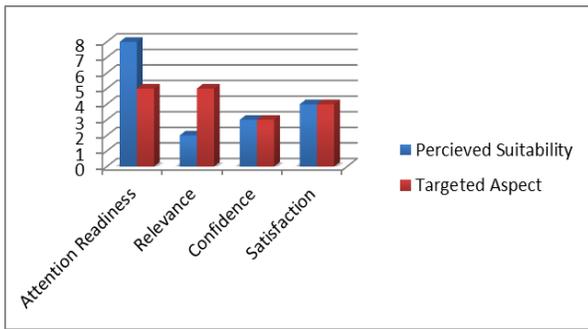
Analysis of the data reveals that among teachers and students, 67 percent of respondents feel they have a high degree of comfort with the use of mobile technology. This finding is significant in supporting the confidence component of Keller’s model and in demonstrating the limited need for instructional time to be invested in building proficiency with the technology. Detracting from the same aspect however, is the apparent lack of comfort amongst teachers with the level of training they have received in developing and implementing mobile learning lessons, where only 50 percent agree they are comfortable and 39 percent respond that they are uncomfortable or highly uncomfortable in this area. Figure1 suggests that although teachers are comfortable with the use of this technology in their classrooms, they are less comfortable in determining how it should be used.

Figure 1. Comfort with technology



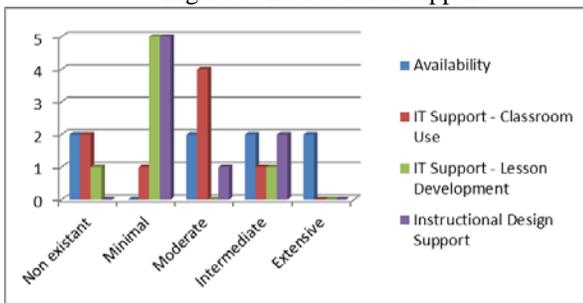
When asked to describe the suitability of mobile learning in supporting Keller’s model for learner motivation and their own intentions for using it in the classroom, teachers give slightly contradictory responses. Figure 2 illustrates a belief that mobile learning is best suited to support the attention readiness and relevance components of learner motivation. Despite this apparent belief, educators admit to targeting each component of ARCS nearly uniformly in lesson design indicating low correlation between perceived suitability and actual use.

Figure 2. Suitability and targeted use of m-Learning



Additionally, educators indicate low levels of support towards the design and implementation of mobile learning. A subsequent segment of the educator survey asked participants to describe the availability of mobile devices for use in the classroom and also the availability of technical and instructional support for those devices; the results are depicted in Figure 3.

Figure 3. Instructional support



The data in this segment are significantly skewed to the left with the mean response for IT support and instructional design being 'minimal' whereas the mean response for availability of technology is 'intermediate.' This pairing may be an indicator of low levels of support for the mobile technologies entering the classroom and as a result, a negative incentive to embed them into the curriculum.

B. Lesson Plan Analysis

The second component of the analysis included a review of a sample of K-12 mobile learning lesson plans. Although representation of each subject area varied, the sample includes at least one lesson plan from each grade level and major subject area. Each lesson plan was reviewed by the research team utilizing a content coding rubric divided into the elements of the ARCS model. Each element was then analyzed within the selected sample of lesson plans. Individual attributes of each lesson plan were scored from (1) - Highly Ineffective through (5) - Highly effective.

- 1) *Attention:* The evaluation team rated the lesson plan sample between effective and highly effectively in their ability to gain and maintain learner attention. Several attributes were evaluated to make this determination as specified below in Table 2.

Table 2. Attention Attributes

	K - 4 th Grade	5 th - 8 th Grade	9 th -12 th Grade
The activity is designed to be	3.7	4.5	4

learner centered.			
The content is likely to stimulate learner curiosity.	4.3	3.1	3.56
The instructional media is likely to stimulate learner curiosity.	4.8	3.2	3.78

Of the sample, K - 4th grade lesson plans scored highest with a mean rating of 4.27, the lowest scoring sample group were the 5th through 8th grade lesson plans scoring an average of 3.60 across the measured attributes.

- 2) *Relevance:* The findings for relevance again demonstrate effective to highly effective incorporation of motivation design theory in lesson planning amongst the sample pool. Again, the Kindergarten through 4th grade lesson plans scored highest across each attribute and overall with a mean score of 4.13. The 5th through 8th grade pool scored the lowest with a mean score of 3.1. The median score of all lesson plans reviewed was 3.67. It is possible that the mobile learning activities are being designed in a way so as to support the millennial learner’s desire to learn content in context and learn through the construction of knowledge.

Table 3. Relevance Attributes

	K - 4 th Grade	5 th - 8 th Grade	9 th -12 th Grade
The instructional media is likely to match the learners preference.	4	3.2	3.67
The application(s) enables the learner to construct knowledge through inquiry.	4.1	2.6	3.56
The iPad application(s) selected is appropriately matched to the learning content.	4.3	3.5	3.89

- 3) *Confidence:* Despite the small sample size and lack of classroom observation of teachers and students impacted by the lesson plans, the content shows a pattern of technology applications apparently appropriately paired to student skill levels with a mean score of 3.90 (Table 4). Conversely, the research team also detected a trend that would suggest learning preference and performance feedback are less emphasized in mobile lesson plan design. The mean score in this dimension for all lesson plans was only 3.52. Despite the limitations of the study, the detection of these patterns may indicate a trend in which educators are thoughtfully selecting applications and exercises matched to the abilities of their students; but the activities selected appear for the most part to be targeted towards one learning style. In this sample, the overall trend is one where learner confidence is least reflected in the lesson plan design.

Table 4. Confidence Attributes

	K - 4 th Grade	5 th - 8 th Grade	9 th -12 th Grade
The iPad application(s) selected is appropriately matched to the abilities of the learner.	3.9	3.6	3.67
The learning activity allows for the emergence of learning styles.	3.3	2.0	3.0

Support and feedback appear to be embedded into the learning activity.	2.3	2.7	2.56
--	-----	-----	------

REFERENCES

- 4) *Satisfaction*: Much like the relevance and confidence components of the lesson plans reviewed, analysis of the learner satisfaction attributes of the content showed varying levels of emphasis across grade levels and characteristics within the satisfaction element. With a mean score of 3.79, the lesson planning seems to support learner satisfaction through the ability for learners to access content in context and exercise freedom of choice in how they learn (Table 5). Conversely, the mean score of 2.9 in the collaboration element indicates that the lesson design may limit collaboration may in turn reduce learner satisfaction.

Table 5. Satisfaction Attributes

The iPad application(s) selected enables learners freedom to explore content.	3	3	3.3
The learning activity is designed to allow for collaboration between learners.	2.8	2.8	3.2
The iPad application(s) selected allows the learner to interact with the learning content in real world context.	4.1	3.7	3.5

C. Recommendations

The following recommendations are suggested to improve mobile learning lesson plan creation: (1) standardize the mLearning lesson plan template to include explicit reference to the state content standard for which the lesson is focused; (2) adopt a uniform structure for lesson learning goals/objectives to include a measurable and observable behavior, condition under which the learning will take occur, and the degree of accuracy for the goal to be considered met by the learner.

If further research were to be conducted in this area in a much larger scale, we recommend the following criteria: (1) a blind sampling of teachers from a diverse group of states and content areas; (2) a larger group of participants who share their feedback, lessons and access for researchers to observe their classrooms; (3) adequate time to study the implementation of mLearning lessons in actual classrooms; (4) Access to students in the classrooms for pre and post-mLearning lesson interviews.

We feel that if more time and resources were allotted to the study of mobile learning implementation in the k-12 learning environment, educators and students would greatly benefit from the findings. According to this study, teachers feel that they have a moderate amount of technology accessible to them for teaching, however, more IT and lesson design support is needed.

ACKNOWLEDGEMENT

This paper is supported by the Oriental Scholar program of Shanghai Municipal Education Commission (TPKY052WMJ).

- [1] E. A. Beckmann, "Leaners on the move: mobile modalities in development studies," *Distance Education*, vol. 31, no.2, pp.159-173, 2010.
- [2] P. Schubert, "Grasping the Realities of Educating in the Digital Age," *Educause Review*, vol. 46, no. 2, pp. 8-9, 2011.
- [3] G. Hwang, and H. Chang, "A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students," *Computers & Education*, vol.56, no.4, pp.1023-1031, 2011.
- [4] A. Kukulska-Hulme, "Learning Cultures on the Move: Where are we heading?" *Journal of educational technology & society*, vol.13, no.4, pp.4-14, 2010.
- [5] M. El-Hussein, and J. C. Cronje, "Defining mobile learning in the higher education landscape," *Journal of educational technology & society*, vol.13, no. 3, pp.12-21, 2010.
- [6] M. McMahon, and R. Popisil. (2005). Laptops for a digital lifestyle: Millennial students and wireless mobile technologies. Available: <http://www.ascilite.org.au/conferences/brisbane05>.
- [7] S. Robert. (2005). Millennial Learning - On Demand Strategies for Generation X and Beyond. Available: <http://www.cedma-europe.org>.
- [8] J. M. Keller, "Development and use of the ARCS model of instructional design," *Journal of instructional development*, vol.10, no.3, pp.2-10, 1987.
- [9] J. L. Shih, H. C. Chu, G. J. Hwang, and Kinshuk "An investigation of attitudes of students and teachers about participating in a context-aware ubiquitous learning activity," *British journal of educational technology*, vol.42, no.3, pp.373-394, 2011.
- [10] M. Warschauer, S. R. Cotten, and M. G. Ames, "One laptop per child Birmingham: Case study of a radical experiment," *International journal of learning and media*, vol.3, no. 2, pp. 61-76, 2011.
- [11] J. Owen, *Program Evaluation: Forms and Approaches*. 3rd Ed. Guildford Press, New York, NY, 2006.

Minjuan Wang is an oriental scholar at Shanghai International Studies University, China, a professor of Learning Design and Technology at San Diego State University, and a Program Manager for the Chancellor's office of California State University. Her research specialties focus on the sociocultural facets of online learning, and the design and development of mobile and intelligent learning. She has published peer-reviewed articles in Educational Technology Research and Development, Computers and Education, Educational Media International, TechTrends, and the British Journal of Educational Technology. She has also published book chapters on engaged learning in online problem solving, Cyberogy for interactive learning online, informal learning via the Internet, and effective learning in multicultural and multilingual classrooms. Address for correspondence: Dr MinjuanWang, 5500 Campanile Dr. PSFA 315, SDSU, San Diego, CA 92182-4561.Tel: 619-5943878 Email:mwang@mail.sdsu.edu.

Yong Chen is an instructional technology specialist at Old Dominion University, Virginia. He got his MA in Educational Technology at San Diego State University. His research interests include online learning, mobile learning, security and privacy in online learning, learning management system safety, and Web 2.0 applications in teaching and learning. He has published 8 papers on peer-review journals, such as the International Review of Research in Open and Distributed Learning, Information and Management, Systems Research and Behavioral Science, and Journal of Information Technology Case and Application Research. Address for correspondence: 335 Gornto, Old Dominion University, Norfolk, VA, 23529. Email: y7chen@odu.edu.

Junli Li is associate professor of Educational Technology, Shanghai International Studies University. She was also a visiting scholar of San Diego State University. Her research focuses on the design and development of mobile learning platforms. Address for correspondence: Dr. Junli Li, Songjiang University Town, SISU, Building 1. Shanghai, China. Email: ljlishisu@163.com.

How to break down the security of an efficient modular exponentiation algorithm

David Tinoco Varela

Universidad Nacional Autónoma de México

Campus Cuautitlán.

Cuautitlán Izcalli, Edo. Mex., México. Email: dativa19@hotmail.com

Abstract—An efficient modular exponentiation algorithm secure against Simple Power Analysis was proposed by Sun Da-Zhi et al. They stated that their algorithm is highly suitable for real-life chip-card applications; however, an attacker can use certain characteristics of the algorithm to obtain the secret key of a cryptosystem. In this paper, we have proposed an attack scheme that obtains the values of three-quarters of the total number of bits of the binary representation of a secret key when the modular exponentiation algorithm proposed by Sun Da-Zhi et al. is executed. This attack uses a combination of Fault attacks and Jacobi symbol attacks to break down the security of the system.

Keywords – Cryptography, modular exponentiation algorithms, side channel attacks, fault attacks, Jacobi symbol.

I. INTRODUCTION

SIDE channel attacks (SCA) are a different form of attack against cryptosystems, first proposed by Kocher [1], who noted that the time consumption or the power traces of an embedded device, when it is executing a cryptographic algorithm, can permit an attacker to obtain the secret key of the cryptosystem by simply observing the signals in a device as an oscilloscope, SCAs are first used to attack the modular exponentiation (or *Add and double* operation in Elliptic Curve Cryptosystems), which is the core operation in cryptosystems such as RSA (*Rivest Shamir Adleman*) [2].

SCAs were the first known physical attacks, but new types of physical attacks appeared subsequently, including the *Fault attack* (FA) proposed by Bonhe, DeMillo and Lipton [3]. FAs are more aggressive than SCAs because FAs physically disturb the execution of the device running the cryptographic algorithm.

The *Square-and-Multiply Always algorithm* was the first algorithm specifically designed to defeat SCAs. This algorithm was proposed by Coron [4], but the algorithm was attacked by the denominated *Safe error attack* [5].

In 2003, Joye and Yen proposed a modular exponentiation algorithm called the *Montgomery powering ladder* to protect cryptosystems against SCAs and FAs [6]. This algorithm works in a regular form, which means that regardless of the value of the bit being processed (0 or 1), the algorithm always will calculate a multiplication followed by a squaring.

The Montgomery powering ladder was modified by Giraud [7] to protect it against FAs by proposing a *Coherence Test* based on a characteristic of the algorithm: the registers in all of the iterations have the form $R_0 = m^x$, $R_1 = m^{x+1}$, and as a result, if the coherence test $R_0 \cdot m = R_1$ is true, then it returns R_0 ; if not, it returns "error".

There are many physical attacks ([8], [9], [10], [11]) that try to break the different modular exponentiation algorithms ([12], [13], [14], [15], [16], [17]), but in 2006, Boreale [18] presented a new type of attack that uses a combination of FA and SCA. According to him, it is possible to obtain the binary string of the secret key d using the *Jacobi symbol* concept. He attacked the *Square and Multiply Right-to-Left* modular exponentiation and proved that his attack is effective even in the presence of message blinding. Schmidt and Medwed [19] used the Jacobi symbol to create an attack that breaks the security of the Montgomery ladder in its blinded form. In the same way, Chong Hee Kim designed an attack in 2010 [20], also based on the Jacobi symbol, to break the security of the *Add-Only* and *Add-Always* algorithms, both algorithms proposed by Joye in 2007 [21].

On the other hand, Sun Da-Zhi et al. provided an efficient algorithm against *Simple Power Analysis* [22], an algorithm that uses two binary strings instead of one to calculate the modular exponentiation.

II. PRELIMINARIES

Some attacks based on the Jacobi symbol have been shown in the literature, and in this paper, a new attack based in the Jacobi symbol will be given. Therefore, the *quadratic residue* concept and the Jacobi symbol concept are explained before the explanation of the attack: for any prime number p , x is a quadratic residue if $\gcd(x, p) = 1$ and $x = y^2 \pmod{p}$ for some y . If $\gcd(x, p) = 1$ but x is not a quadratic residue mod p , then x is called a quadratic non-residue mod p .

$\left(\frac{a}{p}\right)$ is called the *Legendre symbol* of $a \pmod{p}$ and has the following properties:

$$\left(\frac{a}{p}\right) = \begin{cases} 1 & \text{If } a \text{ is a quadratic residue mod } p \\ -1 & \text{If } a \text{ is a quadratic non-residue mod } p \\ 0 & \text{If there is a common factor} \end{cases}$$

Then, we have that $\left(\frac{a}{n}\right) = \left(\frac{a}{p_1}\right) \cdots \left(\frac{a}{p_k}\right)$ is the Jacobi symbol, where $n = p_1 \cdots p_k$, and p 's are prime factors. The Jacobi symbol is a generalization of the Legendre symbol.

Attacks using the Jacobi symbol began with *Boreale*, when he attacked the binary Square and Multiply Right-to-Left modular exponentiation (Algorithm 1) in 2006 [18]. He puts a fault z in R_1 when a squaring is executed in iteration $i - 1$ of the Square-and-Multiply algorithm, and then, depending on the calculation of (S/N) , where S is the output value of the algorithm and N is the modulus, it is possible to know what value of the bit d_i was attacked. This scheme works by assuming that $\left(\frac{m}{N}\right) = 1$, and its behavior is similar to the *Safe error*: if the bit in iteration i is equal to 0, the fault does not affect the calculation of the Jacobi symbol of (R_{0_i}/N) , and z is squaring, so that $(z^2/N) = 1$, but if $d_i = 1$, z affects the register R_{0_i} and can have the Jacobi symbol value $(R_{0_i}/N) = -1$. Then, z can be or not be a quadratic residue. If z is a quadratic residue, then the final result will be $(S/N) = 1$, but if it is a quadratic non-residue, the final result will be $(S/N) = -1$. For this reason, his attack is a probabilistic model.

Algorithm 1 Square and Multiply Right-to-Left.

```

1: Input  $m, d = (d_{n-1}, \dots, d_0)_2$ 
2: Output  $s = m^d$ 
3:  $R_0 \leftarrow 1; R_1 \leftarrow m$ 
4: for 0 to  $n - 1$  do
5:   if  $d_i = 1$  then
6:      $R_0 \leftarrow R_0 \cdot R_1 \bmod N$ 
7:   end if
8:    $R_1 \leftarrow R_1^2 \bmod N$ 
9: end for
10: Return  $R_0$ 

```

After *Boreale*, *Schmidt* [19] proposed an attack consisting of giving a message m with $(m/N) = -1$ to the *Fumaroli-Vigilant algorithm* [23] and skipping the operation $R_{d_i} = R_{d_i}^2$ in the algorithm. Then, by observing the Jacobi symbol of the resulting value, it is possible to learn about the value of d_i and d_{i-1} . If (S/N) is equal to -1, then $d_i = d_{i-1}$.

The attacks mentioned above are easy to implement, and they are powerful because they need to know about only the Jacobi symbol in the returned value by the attacked algorithm to break the security of a cryptosystem.

On the other hand, *Sun Da-Zhi* et al. published a modular exponentiation algorithm that separates the original binary string of the exponent d into two binary strings d_1 and d_2 .

The algorithm is given as algorithm 2, where $sq^{(y)}(A)$ means performing y modular squares on the integer A [22].

Algorithm 2 Algorithm proposed in [22].

```

1: Input  $m, d, N$  with  $d_1 = d_{\lceil k/2 \rceil, 1} \cdots d_{1,1}$  and  $d_2 = d_{\lceil k/2 \rceil, 2} \cdots d_{1,2}$ 
2: Output  $C = m^d \bmod N$ 
3:  $s = m$ 
4:  $C_0 = C_1 = C_2 = C_3 = 1;$ 
5:  $C_{2 \cdot d_{1,2} + d_{1,1}} = s \cdot C_{2 \cdot d_{1,2} + d_{1,1}}$ 
6: for 2 to  $\lceil k/2 \rceil$  do
7:    $s = s \cdot s$ 
8:    $C_{2 \cdot d_{i,2} + d_{i,1}} = s \cdot C_{2 \cdot d_{i,2} + d_{i,1}}$ 
9: end for
10:  $C_1 = C_1 \cdot C_3$ 
11:  $C_2 = C_2 \cdot C_3$ 
12:  $C = sq^{(\lceil k/2 \rceil)}(m^{d_1}) \cdot m^{d_2}$ 

```

The key idea in algorithm 2 is to separate the k -bit binary string d into two $(\lceil k/2 \rceil)$ -bit binary strings d_1 and d_2 , and depending on the values of d_1 and d_2 in each iteration, four registers $C_0 \cdots C_3$ can be utilized to calculate the exponentiation. If $d_{i,1}$ and $d_{i,2}$ (Where $i, 1$ is the i -bit of d_1 and $i, 2$ is the i -bit of d_2) are equal to 0, the chosen register is C_0 ; if $d_{i,1} = 1$ and $d_{i,2} = 0$, the chosen register is C_1 , and so on, for each register.

III. PROPOSED ATTACK AGAINST THE ALGORITHM PROPOSED BY SUN ET AL.

To break the security of the algorithm 2, it is necessary to see its characteristics. The first situation that can be seen is that when $d_{i,1}$ and $d_{i,2}$ are equal to 0, the C_0 value will no longer be occupied to calculate the correct value of the exponentiation: in other words, C_0 is discarded. Therefore, the first step to break down this algorithm is to determine the positions where $d_{i,1} = d_{i,2} = 0$. How can these positions be determined? The answer is simple: the positions can be determined by placing faults in each iteration of the algorithm when it is executing. Depending on whether the output value is correct or incorrect, it is possible to determine whether the bit combination in the attacked position was $d_{i,1} = d_{i,2} = 0$ or any other combination. Clearly, if the bit combination in the attacked position was $d_{i,1} = d_{i,2} = 0$, the output value will be correct. This first step can determine the values of some bits.

It is important to note that under this attack scheme, it will be possible to determine two bits of the original binary string instead of one for each attacked iteration. For example, if the original binary string is equal to $d = 1001010101$, then $d_1 = 10010$ and $d_2 = 10101$, counting from right to left and counting from 0, it is possible to see that in position number three of both d_1 and d_2 , the bits are equal to 0. If the algorithm is running in that position, the register used is

C_0 , and if a fault is placed in that instant, the output result will be correct. Thus, the value of bit number three and the value of bit number eight of the original binary string can be determined. Another perspective is as follows: there are $d = xxxxxxxxxx$, $d_1 = xxxxx$ and $d_2 = xxxxx$, and after the FAs, the sub strings can be seen as $d_1 = x0xxx$, $d_2 = x0xxx$. Therefore, $d = x0xxx0xxx$, where the x 's are the unknown values of the bits of the binary string.

In the second step of the attack, it is necessary to choose a message m such that $(m/N) = 1$ and then to place faults when the algorithm is executing. Because each register is independent in each iteration, and in the last line of the algorithm 2, the only register that is squared is C_1 , it is possible to determine the position where $d_{i,1} = 1$ and $d_{i,2} = 0$ because when only C_1 is attacked, the Jacobi symbol of the output value will always be equal to 1, $(S/N) = 1$, regardless of whether $(z/N) = 1$ or $(z/N) = -1$, where z is the error placed in the calculation. In other words, if an error is placed in C_1 , C_1 will be exponentiated to an even exponent in the last line of the algorithm 2, becoming a quadratic residue, and thus, the Jacobi symbol value of S will be always equal to 1. Taking again the example given in step 1, after the second step has been executed, it is possible to know that $d_{1,1} = 1$ and $d_{1,2} = 0$. From this information, it will be directly known that in the original binary string d , the bit number 1 is equal to 0 and the bit number six is equal to 1. Again, this information can be seen in the next form $d_1 = x0x1x$, $d_2 = x0x0x$, and therefore $d = x0x1x0x0x$.

On average, after step 2, half the values of the original binary string's bits have been found, but there is a characteristic that allows an attacker to know more values of the bits in the binary string d .

The positions where C_0 and C_1 are executed have been determined, and up to here, the positions when C_2 and C_3 are executed remain unknown.

Now, it can be seen that to use the register C_2 , it is necessary that $d_{i,1} = 0$ and $d_{i,2} = 1$, and to use the register C_3 , it is necessary that $d_{i,1} = 1$ and $d_{i,2} = 1$. Then, it can be noted that for both registers, the value of the bit $d_{i,2}$ in i will always be equal to 1, and therefore it is possible to determine that all of the unknown values of the bits in d_2 are equal to 1: thus, $d_1 = x0x1x$, $d_2 = 10101$, and therefore $d = x0x1x10101$. After these three steps, on average, only a quarter of the values of the bits remain unknown.

It is important to note that the binary string d_2 is completely broken; however, the values x in d_1 remain unknown.

To find the unknown bits of the string, it may be possible to use *Partial Key Exposure Attacks* [24]. These attacks permit an attacker to discover the entire secret key while knowing only a fraction of the bits.

IV. EXAMPLE

This section will give an example of the proposed attack over the execution of the algorithm 2. This experiment was realized in a simulated way.

For this example, a random unknown exponent d and a random input value m such that $(m/N) = 1$ were chosen, where

$$m = 523489278217456019834593$$

$$N = 87125894816237589871623581$$

First of all, the correct output value of the modular exponentiation was obtained, which is equal to 84982614287872105537806915. Then, the first step of the attack scheme described above was executed. Table 1 shows the results when faults were placed in the execution of the algorithm:

In table 1, it can be observed that in positions 4 and 6, $d_{i,1} = d_{i,2} = 0$ because there is no error in the output value, and it is possible to determine $d_1 = xx0x0xxxx$, $d_2 = xx0x0xxxx$, and therefore $d = xx0x0xxxxx0x0xxxx$.

The first values of the binary string have been obtained. Now, the next step is to place faults in the execution of the algorithm and to calculate the Jacobi symbol for each output value of the algorithm. Table 2 contains the obtained values of the Jacobi symbol calculation.

The first row in table 2 represents the number of the position that was attacked. It can be observed that in column 0, all values are 0, which is because the calculation of the position 0 in d_1 and d_2 is performed outside the main loop *for* and therefore was not considered in the attack.

It can be noted in table 2 that the Jacobi symbol values in positions 1, 3, 4, 6, y 8 are always equal to 1, but it is known that positions 4 and 6 correspond to C_0 , where $d_{i,1} = 0$ and $d_{i,2} = 0$. Thus, it is possible to know that positions 1, 3 and 8 correspond to C_1 , where $d_{i,1} = 1$ and $d_{i,2} = 0$. With those results, it is feasible to obtain $d_1 = 1x0x01x1x$, $d_2 = 0x0x00x0x$, and therefore $d = 1x0x01x1x0x00x0x$.

Now, it has been learned that $d_{i,2} = 1$ is necessary to execute any of the registers C_2 or C_3 , and thus, it can be determined that all of the unknown bits in d_2 are equal to 1. Thus, $d_2 = 010100101$, and then $d = 1x0x01x1x010100101$, leaving unknown only one quarter, on average, of the values of the binary string d .

The random exponent d remains unknown until the completion of the attack, and after the attack, we know

Attacked position.	Output value.
1	78677014161364257494504919
2	56091741037709499701712765
3	50806030636843748450556941
4	84982614287872105537806915
5	4495242442647294430897836
6	84982614287872105537806915
7	44789352291274716948534142
8	42721333352452789389231267

TABLE 1

Output values when faults were placed in the execution of the algorithm 2.

Number of attack	$i = 0$	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$
Try 1	0	1	-1	1	1	-1	1	-1	1
Try 2	0	1	1	1	1	1	1	-1	1
Try 3	0	1	1	1	1	1	1	1	1
Try 4	0	1	-1	1	1	-1	1	1	1
Try 5	0	1	-1	1	1	1	1	1	1
Try 6	0	1	1	1	1	1	1	-1	1
Try 7	0	1	1	1	1	1	1	-1	1
Try 8	0	1	-1	1	1	1	1	-1	1
Try 9	0	1	1	1	1	1	1	1	1
Try 10	0	1	-1	1	1	-1	1	-1	1
Try 11	0	1	-1	1	1	1	1	1	1

TABLE 2

Jacobi symbol of the output values, when Fault attacks have been placed in the execution of the algorithm 2.

the value of the exponent $d = 153253$. This value can be represented in its binary form as $d = 100101011010100101$. Upon comparing this binary string with the result obtained by the attack, we can note that the found values of the bits are equal to the values in the original string.

V. CONCLUSIONS

We have described a method to break down the modular exponentiation algorithm proposed by Sun Da-Zhi et al. We have observed some characteristics of their algorithm that make it vulnerable to certain attacks, such as FA and Jacobi symbol attacks. The mentioned attacks are used in combination to obtain the values of three-quarters of the total number of bits of the binary representation of an exponent. If we have a binary string of 1024 bits, we can discover 768 bits on average using this scheme, leaving only 256 bits to discover.

We have given an example to demonstrate the effectiveness of our attack.

ACKNOWLEDGMENTS

The authors thank PAPIIT IN112913, PIAPIVC06, and Program PCIC of the UNAM.

REFERENCES

- [1] P. Kocher. "Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems". In *Koblitz, N., ed.: Advances in Cryptology-CRYPTO 96*, 1109 of Lecture Notes in Computer Science:104-113, Springer, 1996.
- [2] RL Rivest, A. Shamir and L. Adleman. "A method for obtaining digital signatures and public-key cryptosystems". *Communications of the ACM*, 21(2):120-126, 1978.
- [3] D. Boneh, R. DeMillo and R. Lipton. "On the importance of checking cryptographic protocols for faults". In *Fumy, W., Ed.: Advances in Cryptology-EUROCRYPT'97. Volume 1233 of Lecture Notes in Computer Science*, pages 37-51, Springer 1997.
- [4] J.S. Coron. "Resistance against differential power analysis for elliptic curve cryptosystems". In *Ko, ., Paar, C., Eds.: Cryptographic Hardware and Embedded Systems-CHES 2002. Volume 1717 of Lecture Notes in Computer Science*, pages 292-302, Springer, 1999.
- [5] S.M. Yen, S. Kim, S. Lim and S. Moon. "A countermeasure against one physical cryptanalysis may benefit another attack". *Information Security and Cryptology-ICISC 2001*, 2288 of Lecture Notes in Computer Science:414-427, Springer, 2001.
- [6] M. Joye and S.M. Yen. "The Montgomery powering ladder". *Cryptographic Hardware and Embedded Systems-CHES 2002*, 2523 of Lecture Notes in Computer Science: 291-302, Springer, 2003.
- [7] C. Giraud. "An RSA implementation resistant to fault attacks and to simple power analysis". *IEEE Transactions on computers*, 55, No 9:1116-1120, IEEE, 2006.
- [8] S.M. Yen, W.C. Lien, S.J. Moon and J.C. Ha. "Power analysis by exploiting chosen message and internal collisions—vulnerability of checking mechanism for RSA-decryption". *Progress in Cryptology-Mycrypt 2005*, 3715 of Lecture Notes in Computer Science:183-195, Springer, 2005.
- [9] C.H. Kim and J.J. Quisquater. Fault attacks for CRT based RSA: New attacks, new results, and new countermeasures. *Information Security Theory and Practices. Smart Cards, Mobile and Ubiquitous Computing Systems*, 4462:215-228, Springer, 2007.
- [10] S. Chari, J. Rao and P. Rohatgi. "Template attacks". *Cryptographic Hardware and Embedded Systems-CHES 2002*, 2523 of Lecture Notes in Computer Science:12-28, Springer 2002.
- [11] S.M. Yen and M. Joye. "Checking before output may not be enough against fault-based cryptanalysis". *IEEE Transactions on Computers*, 49(9):967-970, 2000.
- [12] H. Mamiya, A. Miyaji and H. Morimoto. "Efficient countermeasures against RPA, DPA, and SPA". *Cryptographic Hardware and Embedded Systems-CHES 2004*, 3156 of Lecture Notes in Computer Science:343-356, Springer, 2004.
- [13] C.C. Lu, S.Y. Tseng and S.K. Huang. "A secure modular exponential

- algorithm resists to power, timing, C safe error and M safe error attacks". In *19th International Conference on Advanced Information Networking and Applications, 2005. AINA 2005*, volume 2, pages 151-154, IEEE, 2005.
- [14] C.H. Kim and J.J. "Quisquater. How can we overcome both side channel analysis and fault attacks on RSA-CRT?". *Workshop on Fault Diagnosis and Tolerance in Cryptography*, pages 21-29, IEEE, 2007.
- [15] A. Boscher, R. Naciri and E. Prouff."CRT RSA algorithm protected against fault attacks". *Information Security Theory and Practices. Smart Cards, Mobile and Ubiquitous Computing Systems*, 4462 of LNCS: 229-243, Springer, 2007.
- [16] J.C. Ha, C.H. Jun, J.H. Park, S.J. Moon and C.K. Kim. "A New CRT-RSA Scheme Resistant to Power Analysis and Fault Attacks". *Third 2008 International Conference on Convergence and Hybrid Information Technology*:351-356, IEEE, 2008.
- [17] A. Boscher, H. Handschuh and E. Trichina. "Blinded fault resistant exponentiation revisited". In *L. Breveglieri, S. Gueron, I. Koren, D. Naccache, and J.-P. Seifert, editors, Workshop on Fault Diagnosis and Tolerance in Cryptography - FDTC'09*, pages 3-9, IEEE, 2009.
- [18] M. Boreale. "Attacking right-to-left modular exponentiation with timely random faults". *Fault Diagnosis and Tolerance in Cryptography*, 4236 of LNCS: 24-35, Springer, 2006.
- [19] Jörn-Marc Schmidt and Marcel Medwed. "Fault Attacks on the Montgomery Powering Ladder". *13th Annual International Conference on Information Security and Cryptology, Proceedings*, LNCS, Springer, 2010.
- [20] C.H. Kim. "New fault attacks using Jacobi symbol and application to regular right-to-left algorithms". *Information Processing Letters*, 110(20):882-886, Elsevier, 2010.
- [21] M. Joye."Highly regular right-to-left algorithms for scalar multiplication". *Cryptographic Hardware and Embedded Systems-CHES 2007*, 4727 of Lecture Notes in Computer Science:135-147, Springer, 2007.
- [22] D.Z. Sun, J.P. Huai, J.Z. Sun and Z.F. Cao. "An efficient modular exponentiation algorithm against simple power analysis attacks". *Consumer Electronics, IEEE Transactions on*, 53(4):1718-1723, 2007.
- [23] G. Fumaroli and D. Vigilant. "Blinded fault resistant exponentiation". *Fault Diagnosis and Tolerance in Cryptography*, 4236 of Lecture Notes in Computer Science:62-70, Springer-Verlag 2006.
- [24] Johannes Blömer and Alexander May. "New partial key exposure attacks on RSA". *Advances in Cryptology-CRYPTO 200*, 3:27-43, Springer 2003.

Objective Stimulus Features for Predicting Human Judgments of Visual Pattern Goodness: An Empirical Comparison

Godfried T. Toussaint

Abstract—Pattern goodness is a concept that attempts to capture the perceptual salience of a pattern. One of its main features is the property of global mirror symmetry, which has received a great deal of attention in the literature. Here global mirror symmetries are empirically compared with local symmetries and other geometric objective stimulus features, in terms of how well they predict human ratings of pattern goodness, using a well-known dataset comprised of 17 dot patterns consisting of five dots arranged on a 3×3 grid. The main conclusion is that for small patterns the global symmetries outperform local symmetries and other geometric features.

Keywords—Visual pattern goodness, pattern complexity, local and global symmetries, palindromes, compactness, uniformity, geometric stimulus features.

I. INTRODUCTION

THIS paper presents an empirical comparative evaluation of several geometric properties of stimulus patterns, to determine how well they correlate with human judgments of pattern goodness. The word pattern has many different meanings in disparate fields of knowledge [1]. For the purpose of the present investigation, a pattern is considered to be a visual two-dimensional geometrical form, figure, or shape. Examples of such patterns include dot-patterns [2], [3] and polygonal shapes [4]-[7]. The concept of *pattern goodness* is somewhat more difficult to define precisely. Following the tradition of the first Gestalt psychologists, Nucci and Wagemans define the goodness of a pattern, in very general terms, as its salience or perceptual strength [8]. Similarly Wendell Garner considers good patterns to be those that have a lot of pattern-ness, and poor patterns those that have little pattern-ness [9]. On the other hand Gordon Bear considers sets of elements to be “good” patterns provided that they are perceived as simple and well organized, and “poor” if they are perceived as grouped into complex disorganized patterns [10]. Thus Bear suggests that pattern goodness might be ordered along a simplicity-complexity continuum. Indeed, Lane and Evans have found empirical evidence that ratings of pattern goodness and pattern complexity are inversely related [11].

This work was supported by a grant from the Provost’s Office through the Faculty of Science at New York University Abu Dhabi.

G. T. Toussaint is Professor of Computer Science, and the Head of the Computer Science Program at New York University Abu Dhabi, in Abu Dhabi, United Arab Emirates (corresponding author e-mail: gt42@nyu.edu).

An objective computational modeling approach to pattern goodness seeks to determine which objective stimulus features correlate with human judgments of the goodness of a given pattern. One objective feature that has played a pivotal role in the exploration of operational models of pattern goodness is global mirror symmetry [12]-[14]. Furthermore, Hamada and Ishihara found empirically that human judgments of pattern complexity are inversely related to the amount of symmetry present in a pattern [3]. However, global mirror symmetry is a binary all-or-nothing property, and patterns in the real world usually exhibit only partial or approximate mirror symmetries that lie somewhere in between these two extremes. For this reason some authors have proposed methods that quantify the amount of different symmetries possessed by a pattern. For example, van der Helm and Leeuwenberg suggested a graded property of symmetry they dubbed holographic regularity [15]. A completely different approach was taken by Zabrodsky, Peleg, and Avnir, who proposed to measure the amount of symmetry in a pattern P by the similarity (or the inverse of distance in some suitable metric space) between P and the nearest symmetric pattern among a family of symmetric patterns [16]. However, both of these approaches entail rather lengthy and complex calculations.

In 1968 Alexander and Carey proposed a surprisingly simple objective measure of hierarchical symmetry for the case of visual one-dimensional binary patterns [17]. Their measure just counts the total number of subsymmetries, i.e., contiguous palindromic subsequences that are present in a pattern. This method is simple to calculate by hand for short sequences, and may also be computed efficiently for long sequences [18], [19]. Using a dataset of 35 binary sequences consisting of black and white squares, Alexander and Carey performed several tests with human subjects to determine the perceived complexity of the sequences. The main conclusion of their paper is that patterns that contain many subsymmetries are cognitively simple, whereas patterns that contain few subsymmetries are not cognitively simple. They report that the correlation between the rankings of the 35 patterns by the number of subsymmetries and the human judgments is 0.808 with a significance level better than 0.00001. It is worth noting that calculation of the Spearman rank correlation coefficient using the *VassarStats: Website for Statistical Computation*

(<http://www.vassarstats.net>) yields a higher correlation of 0.867 with a significance level better than 0.000001. This impressive performance by the subsymmetry measure in the domain of visual pattern perception motivated its exploration in the domain of auditory temporal musical rhythmic patterns; it was found for several datasets that the Spearman rank correlations fell in the range between 0.662 and 0.719, with significance levels better than 0.001 in the worst case, and that the subsymmetry measure performed even better than the *Syncoption Index* of Longuet-Higgins and Lee [20], [21].

In a series of seven experiments to investigate the objective stimulus features that determine judgments of perceived pattern complexity, Chipman proposed and explored an extension of the Alexander-Carey subsymmetry measure (referred to as partial or approximate symmetry) to two-dimensional patterns, as well as its generalization to variable weighting of symmetries of differing lengths [22]. The dataset was composed of 45 binary patterns consisting of 6×6 matrices of black and white squares. A multiple linear regression analysis of the results showed that the sum of the vertical and horizontal subsymmetries yielded a significantly high correlation of 0.72 with human judgments, outperforming a list of several other structural features of the black area. The patterns in the Chipman data are large enough and sufficiently complicated so that only a small fraction of the 45 patterns contain global (exact all-or-nothing) mirror or rotational symmetries [22]. This is also evident in the Alexander-Carey one-dimensional data, for which subsymmetries performed so well [17]. Thus for these datasets the symmetries calculated are in effect predominantly local subsymmetries. If patterns contain an insignificant number of global symmetries, then measures of local subsymmetries are likely to outperform measures of global symmetries. Furthermore, subsymmetries tacitly encode the hierarchical structural information contained in a pattern, and such information is considered relevant to the perception of pattern goodness, complexity, and symmetry [15], [23]. This raises the question of scalability with respect to the size of the patterns tested: are subsymmetries still useful for small patterns that contain global symmetries?

In this paper global mirror symmetries are empirically compared with local symmetries and several new geometric objective stimulus pattern features previously unexplored, in terms of how well they predict human judgments of pattern goodness, using a well-known dataset first studied by Garner and Clement [24], and subsequently explored by several other researchers [9], [10], [25]-[29]. The dataset is comprised of 17 dot patterns consisting of five dots arranged in an imaginary 3×3 matrix as shown in Fig. 2. These patterns are small enough to possess a range of global mirror and rotational symmetries, and thus test the efficacy of local subsymmetries as compared to global symmetries. Although the role of global symmetries has been studied extensively as a determinant of pattern goodness, almost no work has been done comparing the interaction of global and local symmetries [8].

II. SUBSYMMETRIES

A. One-Dimensional Patterns

The objective geometric method proposed by Alexander and Carey to measure the complexity of one-dimensional sequences of black and white squares simply counts the total number of subsymmetries present in the sequence [17]. A subsymmetry is a contiguous subsequence that has mirror symmetry, i.e., is palindromic. The example in Fig. 1 illustrates the process with two sequences of length six shown at the top. The sequence on the left has four subsymmetries of length 3, and two of length 5, for a total of six. The sequence on the right has two subsymmetries of length 2, and two of length 3, for a total of four. A pattern with relatively many subsymmetries is considered to be the simpler than one with relatively few subsymmetries. Therefore by this measure the sequence on the left is considered simpler than the one on the right. Note that since neither pattern exhibits global mirror symmetry, the presence or absence of global symmetry fails to differentiate these two patterns.

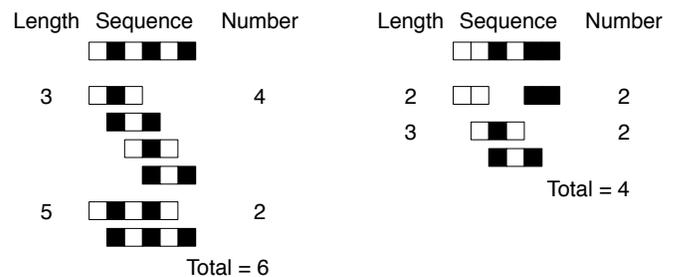


Fig. 1 Example of calculation of subsymmetries.

B. Two-Dimensional Patterns

Unlike one-dimensional patterns, two-dimensional patterns defined in terms of binary elements in square matrices may exhibit mirror symmetries not only along vertical axes, but also about the horizontal and diagonal axes. Furthermore they may exhibit rotational symmetries with angles that are multiples of 90°. Consider the three patterns in Fig. 2 taken from the Garner-Clement dataset shown in Fig. 4. These three patterns (numbered 1, 3, and 15 in Fig. 4) are shown embedded in their 3×3 generating matrices for the sake of clarity. However, in the experiments performed by Garner and Clement the patterns were also presented to subjects without these grids, as depicted in Fig. 4. They concluded that the average rating of the pattern goodness did not vary significantly with either of these two methods of presentation. Pattern (a) has four mirror symmetries along the vertical, horizontal, and positive and negative slope diagonal axes, as well as four rotational symmetries (90°, 180°, 270°, 360°), for a total of 8. Pattern (b) has one mirror symmetry along the vertical axis, and one rotational symmetry (360°), for a total of 2. Pattern (c) contains no mirror symmetries, and one rotational symmetry (360°).

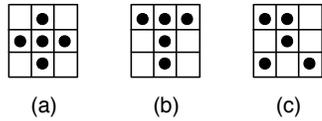


Fig. 2 Types of symmetries in 2-dimensional patterns.

There are two natural ways of generalizing the calculation of subsymmetries from one dimension to two, illustrated in Fig. 3. Consider the pattern shown at the top of Fig. 3 (this is pattern number 13 in Fig. 4). The first approach measures linear (1-dimensional) subsymmetries, as was originally done by Alexander and Carey [17] with 1-dimensional patterns, but now in four different directions: horizontal, vertical positive diagonal, and negative diagonal, shown, respectively in Fig. 3 (a), (b), (c), and (d). The final value obtained is the sum of all the subsymmetries in the four directions. Furthermore, the calculations may be done by either including or excluding the empty squares as part of the pattern. Thus if the empty squares are taken into account the subsymmetries for (a), (b), (c), (d) are, respectively, 5, 3, 2, 2, for a total of 12, whereas if subsymmetries are calculated only with the dots, one obtains, respectively 3, 2, 1, 1, for a total of 7. The second approach to measuring subsymmetries in two-dimensional (2-D) patterns is by counting the number of 2-D subsymmetries present in contiguous 2-D subsets of the patterns. First the 3×3 pattern may be divided into four 2×2 sub-patterns as illustrated in Fig. 3 (e), (f), (g), (h). Then each 2×2 sub-pattern may be examined not just for its mirror symmetries, but for the rotational symmetries as well. Thus, for the pattern in Fig. 3 the 2×2 sub-patterns (e), (f), (g), (h) yield, respectively, 1, 1, 2, 1 mirror symmetries (horizontal, vertical, and \pm diagonals), as well as 1, 1, 2, 1 rotational symmetries. Both of these methods are compared below with global symmetries using the Garner and Clement dataset [24] shown in Fig. 4.

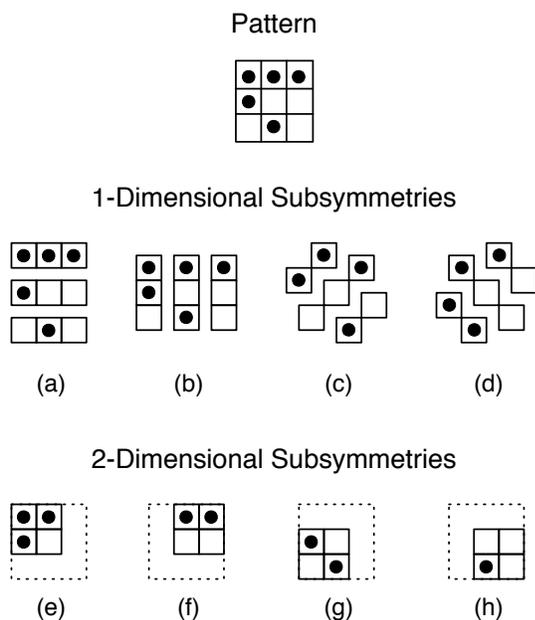


Fig. 3 Two methods of measuring subsymmetries in 2-D.

III. THE DOT-PATTERN DATASET

A. The Dot Pattern Dataset of Garner and Clement

Garner and Clement [24] designed an experiment to test the hypothesis that judgments of the goodness of a pattern P are inversely related to the size of a psychologically inferred set of patterns equivalent to P , under rotation and mirror image transformations. To test their hypothesis they designed a dataset consisting of 90 patterns, each made up of precisely five dots arranged on a 3×3 matrix, with the additional constraint that every column and row should contain at least one dot. The 90 patterns consisted of mirror images and rotations of the 17 patterns shown in Fig. 4. Two experiments were performed. In the first experiment the subjects had to rate each of the 90 patterns on a 7-point scale of goodness, 1 being the best and 7 the worst. The mean rating across all subjects is shown in the third column from the left in Fig. 4. In the second experiment subjects were told to group the patterns into approximately eight groups according to similarity. The score for each pattern was the size of the group in which it was placed, and represented an estimate of the size of the psychologically inferred set. The mean group size across all subjects is shown in the fourth column from the left in Fig. 4. The authors report that the linear correlation between these two variables was 0.84 for the dot patterns presented without their encasing 3×3 grids, thus confirming their hypothesis. The Spearman rank correlation between these two rankings is 0.875 at a significance level of 0.000002. On the other hand, rather than using a subjective measure of the psychologically inferred subset size for each pattern, one can measure an objective mathematical stimulus feature such as the number of global reflection and rotation symmetries possessed by each pattern. These numbers are listed in the third column from the right in Fig. 4 under the title Global Sym. Garner and Clement [24] argue that uncertainty is the fundamental factor in pattern goodness, and that these symmetries are merely attendant properties of uncertainty. On the other hand the Spearman rank correlation between the average ratings and global symmetries in Fig. 4 is -0.889 with a p -value of 0.000001, outperforming the subjective measure of uncertainty based on the size of inferred subsets.

IV. RESULTS

A. Global Versus Local Symmetries

The first main question addressed in this study is whether local subsymmetries of patterns are factors that contribute to the determination of goodness and complexity ratings when patterns possess relatively many global symmetries. Previous studies showed that subsymmetries correlated highly with measures of pattern goodness and pattern complexity in the cases of 1-dimensional patterns of length 7 [17], and 2-dimensional patterns defined on 6×6 matrices [22]. However, the patterns in these datasets contain almost no global symmetries. Therefore the Garner-Clement dataset of Fig. 4 provides a litmus test for exploring this question since it contains patterns that exhibit a range of global symmetries.

ID No.	Pattern	Dot Pattern Data Set				
		Mean Rating	Group Size	Global Sym	Local Sym	Dot Local Sym
1		1.00	9.35	8	16	10
2		1.03	8.25	8	16	10
3		1.55	9.80	2	12	8
4		1.71	12.69	2	10	6
5		1.74	14.04	2	12	8
6		1.78	11.26	2	14	8
7		1.77	10.40	2	12	8
8		2.24	12.16	2	12	8
9		3.05	15.36	2	14	8
10		3.50	16.69	2	10	6
11		3.40	14.52	1	14	9
12		4.59	15.43	1	12	8
13		4.77	16.81	1	12	7
14		4.80	16.37	1	12	7
15		5.19	16.39	1	12	8
16		5.11	16.69	1	12	7
17		5.49	15.74	1	10	7

Fig. 4 The dot pattern dataset of Garner and Clement.

The local symmetries (subsymmetries) were first calculated using the linear 1-dimensional method in all four directions (vertical, horizontal, and diagonal) as illustrated in Fig. 3 (a) – (d). The sums of all the subsymmetries in all four directions for each pattern are shown in the second column from the right in Fig. 4 under the heading Local Sym. The Spearman rank correlation between the Local Sym values and the subjective mean ratings is -0.436 , significant at the 0.04 level. This is appreciably worse than the -0.889 obtained with global symmetries. However, it is reasonable to hypothesize that symmetries among the empty white boxes in the grid representation of the patterns (as depicted in Fig. 3) should perhaps not be counted, since the subjects were not shown these grids in this test, and one might expect the foreground black dots to play a more dominant role in the perception of the pattern subsymmetries than the empty background white regions. Therefore the calculation was repeated using only the black dots. The results of this calculation are shown in rightmost column of Fig. 4 under the heading Dot Local Sym. The Spearman rank correlation between the Dot Local Sym values and subjective mean ratings is -0.512 significant at the 0.017 level. This is somewhat better than the -0.436 , suggesting that the black dot symmetries are somewhat more salient than the symmetries present in the empty squares. Nevertheless, this result still falls quite short of the -0.889 obtained with global symmetries. This suggests that for small patterns that contain a large number of global symmetries, the global symmetries are more salient than the local symmetries for the determination of pattern goodness, when the local symmetries are calculated in this linear fashion.

The local symmetries were also calculated in the 2-dimensional mode as illustrated in Fig. 3 (e) – (h). Recall that in this mode, when examining 2×2 sub-patterns of the 3×3 pattern, in addition to calculating the rotational symmetries, the mirror symmetries in all four directions (vertical, horizontal, and diagonals) may also be computed. The total number of 2×2 subsymmetries did not correlate at all with the mean ratings ($r = -0.047$). It also did not correlate with the global symmetries ($r = 0.185$). Furthermore, adding the number of global and local symmetries together resulted in a correlation of -0.395 with a p value of 0.058. If the subsymmetries are calculated with only the black dots (ignoring the empty squares) then this correlation rises to -0.561 significant to the 0.01 level. Therefore incorporating the subsymmetries in this 2-dimensional fashion also brings down the -0.889 correlation of the global symmetries considerably.

B. Geometric Pattern Features

Besides global and local symmetries, some of the patterns in the Garner-Clement dataset of Fig. 4, contain geometric properties that would appear to contribute to pattern goodness somewhat independently of global and local symmetries. One of these features is the presence of three consecutive collinear dots (collinear triplets), also called *straight lines* by Garner and Clement [24]. For example, patterns 6, 8, and 10 all

contain 2 global symmetries, but pattern 6 contains two collinear triplets (one horizontal and one vertical), pattern 8 contains one diagonal triplet, and pattern 10 contains none, suggesting that this is the reason that subjects rated patterns 6, 8, and 10 as being progressively poorer (mean ratings of 1.78, 2.24, and 3.5, respectively). Garner and Clement observed this behavior, but concluded from an analysis of variance, that the number of collinear triplets is a much less important factor than the number of global symmetries [24]. However, if the number of collinear triplets is compared independently with the mean ratings, the correlation is -0.595 significant at the 0.005 level. Therefore this geometric property, although not quite as strong a predictor as the number of global symmetries, is nevertheless quite salient on its own. A natural question is whether the success of the number of triplets is due to the fact that it may be implicitly taking global symmetries into account. However, these two variables are not significantly correlated with each other. The correlation computed for the rankings by number of global symmetries and number of collinear triplets is 0.364 with a significance value of 0.08 .

Several researchers have classified objective features of pattern goodness into three broad categories: *uniformity*, *compactness*, and *symmetry* [30]-[32]. Uniformity may be quantified by the lack of variety of different elements in the pattern. The collinear triplets discussed above fall under this umbrella: a higher number of collinear triplets indicates more uniformity. Typically such measures have been applied to geometric patterns such as polygons [7], for which the features tested are the number of edges, intersections of edges, and angles between edges; but such features are not present in the dot patterns of Fig. 4. One way to enable dot patterns to comply with such measurements is to construct geometric proximity graphs of the dot patterns [33]-[38]. These graphs serve as perceptual primal sketches of the dot pattern, and also facilitate the definitions of objective measures of uniformity. One proximity graph that has been particularly successful at capturing the perceptual structure of a dot pattern is the *relative neighborhood graph* (RNG) [33]-[35]. The RNG is obtained by adding an edge between two dots if they are closer to each other than to any other dot. Fig. 5 shows the RNG of patterns numbered (in left to right order) 1, 7, and 15 in Fig. 4. One measure of uniformity of a dot pattern is the number of distinct directions of the edges of its RNG. Thus the RNGs in Fig. 5 have (in left to right order) 2, 3, and 4 distinct directions, respectively. The correlation between the ranking by this variable and the mean ratings is 0.587 significant at the 0.007 level, similar in terms of predictive power to the number of collinear triplets ($r = 0.595$, $p = 0.005$).

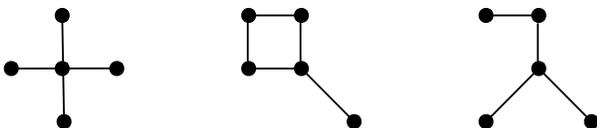


Fig. 5 The relative neighborhood graphs of three dot patterns.

Compactness on the other hand may be quantified by Gestalt measures of proximity. Perhaps the simplest and most natural measure of compactness is the sum of the distances of the pattern elements (in this case the dots) from the center of the pattern (in this case the center of the 3×3 grid in which the dot patterns are embedded). Perhaps not surprisingly, since this measure of compactness ignores structural aspects such as symmetry, it does not correlate significantly with the mean goodness ratings ($r = 0.235$ with $p = 0.18$).

C. Combining Global Symmetries with Collinear Triplets

Since the number of global symmetries and the number of collinear triplets contained in a pattern are not significantly correlated with each other, but each feature is quite salient on its own for predicting the mean goodness ratings, one is naturally tempted to combine these two features according to a weighted linear sum. However, this raises the question of how much the triplets should be weighted relative to the global symmetries. Simply adding the two variables together yields a correlation of -0.869 , significant at the 0.000003 level, which is in fact a fraction worse than the global symmetries applied in isolation (-0.889). Therefore, this choice of weights places too much emphasis on the number of triplets. On the other hand, if the number of symmetries contained in a pattern that has one collinear triplet is increased by 0.3 , and by 0.7 for one that has two collinear triplets, then the correlation increases markedly to -0.935 , significant at the 0.000001 level.

V. CONCLUSION

Several researchers in the past have compared the saliency of a variety of measures of symmetry, uniformity, and compactness, in terms of predictability of pattern goodness. From experiments with polygonal shapes, Marković concludes that symmetry is a significant constraint of pattern goodness, whereas uniformity and compactness fail in this regard [30]. Here, contrary to Marković's conclusion, for the case of small dot patterns, a new proposed objective measure of uniformity, defined as the number of different orientations of the edges in the relative neighborhood graph (RNG) of the dot pattern, is shown to correlate highly with human judgments of pattern goodness. It is also shown that a natural measure of compactness, the sum of distances from the center of a pattern to all the dots, fails as a predictor of pattern goodness. According to Nucci and Wagemans [8] all measures of pattern goodness single out symmetry as their most salient component, but the study of the distinction and relation between local and global symmetries has been neglected. Here global symmetries are compared not only to geometric features of uniformity and compactness, but with a type of local symmetries called subsymmetries (sub-palindromes). The main conclusion is that for small patterns such as those in Fig. 4, that contain many global symmetries, global symmetries are much better predictors of pattern goodness than any other measure tested, and local symmetries in fact degrade the saliency of global symmetry measures when both are combined. Finally, the best overall result is obtained with a combination of global symmetries and the number of collinear triplets present in the patterns ($r = -0.935$ with $p = 0.000001$).

REFERENCES

- [1] E. R. Toussaint, and G. T. Toussaint, "What is a pattern?" *Proceedings of Bridges: Mathematics, Music, Art, Architecture, Culture*, Gwacheon National Science Museum, Seoul, Korea, August 14-19, 2014, pp. 293-300.
- [2] N. Ahuja, and M. Tuseryan, "Extraction of early perceptual structure in dot patterns: Integrating region, boundary, and component Gestalt," *Computer Vision, Graphics, and Image Processing*, vol. 48, pp. 304-356, 1989.
- [3] J. Hamada, and T. Ishihara, "Complexity and goodness of dot patterns varying in symmetry," *Psychological Research*, vol. 50, pp. 155-161, 1988.
- [4] J. J. Gibson, "What is a form?" *Psychological Review*, vol. 58, pp. 403-412, 1951.
- [5] B. Pinna, "What is the meaning of shape?" *Gestalt Theory*, vol. 33, no. 3/4, pp. 383-422, 2011.
- [6] J. Zunic, and P. L. Rosin, "A new convexity measure for polygons," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 923-934, 2004.
- [7] F. Attneave, "Physical determinants of the judged complexity of shapes," *Journal of Experimental Psychology*, vol. 53, no. 4, pp. 221-227, 1957.
- [8] M. Nucci, and J. Wagemans, "Goodness of regularity in dot patterns: global symmetry, local symmetry, and their interactions," *Perception*, vol. 36, pp. 1305-1319, 2007.
- [9] W. R. Garner, "Good patterns have few alternatives: Information theory's concept of redundancy helps in understanding the concept of goodness," *American Scientist*, vol. 58, no. 1, pp. 34-42, January-February 1970.
- [10] G. Bear, "Figural goodness and the predictability of figural elements," *Perception and Psychophysics*, vol. 13, pp. 32-40, 1973.
- [11] S. H. Lane, and S. H. Evans, "Judged complexity as a function of schema related variables," *Psychonomic Science*, vol. 11, pp. 45-46, 1968.
- [12] P. Locher, and C. Nodine, "The perceptual value of symmetry," *Computers & Mathematics with Applications*, vol. 17, pp. 475-484, 1989.
- [13] J. Saarinen, "Detection of mirror symmetry in random-dot patterns at different eccentricities," *Vision Research*, vol. 28, pp. 755-759, 1988.
- [14] H. B. Barlow, and B. C. Reeves, "The versatility and absolute efficiency of detecting mirror symmetry in random dot displays," *Vision Research*, vol. 19, pp. 783-793, 1979.
- [15] P. A. van der Helm, and E. L. J. Leeuwenberg, "Goodness of visual regularities: A nontransformational approach," *Psychological Review*, vol. 103, no. 3, pp. 429-456, 1996.
- [16] H. Zabrodsky, S. Peleg, and D. Avnir, "A measure of symmetry based on shape similarity," *Proceedings of the Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Champaign, IL, June 15-18, 1992, pp. 703-706.
- [17] Alexander, C. & Carey, S. (1968) Subsymmetries, *Perception & Psychophysics*, 4, pp. 73-77.
- [18] S. J. Pan, and R. C. T. Lee, "Looking for all palindromes in a string," *Proceedings of the 23rd Workshop on Combinatorial Mathematics and Computation Theory*, Da-Yeh University, Changhua, Taiwan, April 28-29, 2006, pp. 166-170.
- [19] R. Groult, E. Prieur, and G. Richomme, "Counting distinct palindromes in a word in linear time," *Information Processing Letters*, vol. 110, pp. 908-912, 2010.
- [20] G. T. Toussaint, and J. F. Beltran, "Subsymmetries predict auditory and visual pattern complexity," *Perception*, vol. 42, pp. 1095-1100, November 2013.
- [21] H. C. Longuet-Higgins, and C. S. Lee, "The rhythmic interpretation of monophonic music," *Music Perception*, vol. 1, pp. 424-441, 1984.
- [22] S. F. Chipman, "Complexity and structure in visual patterns," *Journal of Experimental Psychology: General*, vol. 106, no.3, pp. 269-301, 1977.
- [23] H. Zabrodsky, S. Peleg, and D. Avnir, "Hierarchical symmetry," *Proceedings of the International Conference on Pattern Recognition*, The Hague, 1992, pp. 9-12.
- [24] W. R. Garner, and D. E. Clement, "Goodness of pattern and pattern uncertainty," *Journal of Verbal Learning and Verbal Behavior*, vol. 2, pp. 446-452, 1963.
- [25] D. E. Clement, "Uncertainty and latency of verbal naming responses as correlates of pattern goodness," *Journal of Verbal Learning and Verbal Behavior*, vol. 3, pp. 150-157, 1964.
- [26] S. Handel, and W. R. Garner, "The structure of visual pattern associates and pattern goodness," *Perception and Psychophysics*, vol. 1, pp. 33-38, 1966.
- [27] D. E. Clement, and F. Sistrunk, "Judgments of pattern goodness and pattern preference as functions of age and pattern uncertainty," *Developmental Psychology*, vol. 5, no. 3, pp. 389-394, 1971.
- [28] R. J. Glushko, "Pattern goodness and redundancy revisited: Multidimensional scaling and hierarchical clustering analyses," *Perception and Psychophysics*, vol. 17, no. 2, pp. 158-162, 1975.
- [29] J. Takahashi, Y. Kawachi, and J. Gyoba, "Internal criteria underlying affective responses to visual patterns," *Gestalt Theory*, vol. 34, no. 1, pp. 67-80, 2012.
- [30] S. Marković, "Objective constraints of figural goodness," *Psihologija*, vol. 35, br.3-4, pp. 245-260, 2002.
- [31] M. Krüger, "Binary sequences. II. Homogeneity and symmetry," *Information Sciences*, vol. 31, pp. 15-31, 1983.
- [32] F. Papentin, and M. Krüger, "Binary sequences. III. Complexity versus homogeneity and symmetry," *Information Sciences*, vol. 31, pp. 33-39, 1983.
- [33] G. T. Toussaint, "The relative neighborhood graph of a finite planar set," *Pattern Recognition*, vol. 12, pp. 261-268, 1980.
- [34] G. T. Toussaint, "Applications of the relative neighborhood graph," *International Journal of Advances in Computer Science & Its Applications*, vol. 4, issue 3, pp. 77-85, 2014.
- [35] J. W. Jaromczyk, and G. T. Toussaint, "Relative neighborhood graphs and their relatives," *Proceedings of the IEEE*, vol. 80, no. 9, pp. 1502-1517, 1992.
- [36] G. T. Toussaint, "A graph-theoretical primal sketch," in *Computational Morphology*, G. T. Toussaint, Ed., North-Holland, pp. 229-260, 1988.
- [37] D. Avis, and J. Horton, "Remarks on the sphere of influence graph," *Discrete Geometry and Convexity* (New York, 1982), pp. 323-327, *Annals New York Academy of Science*, 440, New York, 1985.
- [38] G. T. Toussaint, "The sphere-of-influence graph: Theory and applications," *Proceedings of the 3rd International Conference on Information Technology, System, & Management*, Abu Dhabi, United Arab Emirates, pp. 43-47, May 8-9, 2014.

Godfried T. Toussaint is a Professor and head of Computer Science at New York University Abu Dhabi, in the United Arab Emirates. He is also an affiliate researcher in the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology in Cambridge, MA, USA. For many years he taught and did research in the School of Computer Science at McGill University in Montreal, in the areas of information theory, pattern recognition, textile-pattern analysis and design, computational geometry, machine learning, music information retrieval, and computational music theory. In 2005 he became a researcher in the Centre for Interdisciplinary Research in Music Media and Technology, in the Schulich School of Music at McGill University. He is a founder and co-founder of several annual international conferences and workshops, including the *ACM Symposium on Computational Geometry*, and the *Canadian Conference on Computational Geometry*. He is an editor of several journals, including *Computational Geometry: Theory and Applications*, the *International Journal of Computational Geometry and Applications*, *ISRN Geometry*, and the *Journal of Mathematics and the Arts*. He received several distinguished awards including a *Killam Fellowship* from the Canada Council for the Arts, and in 2009 a *Radcliffe Fellowship* from Harvard University, where he spent one year at the Radcliffe Institute for Advanced Study, and one year in the Music Department. His research on the phylogenetic analysis of musical rhythms has been reported in several media, and was the focus of two Canadian television programs.

Extending Cloud Computing and Learning for Mobility

Phil Robisch, Rebecca J. Kirsininkas, and Minjuan Wang

Abstract— This paper addresses how mobile learning extends the accessibility and personalization of learning delivered through cloud services and use of mobile devices. The theory of mobile cloud learning is explained comparing the advantages and disadvantages. Three primary learning systems critical to cloud learning are presented with examples of applications and their uses – Learning Management Systems (LMS), Student Management Systems, and Video Capture with Web Conferencing Systems. Additionally, two separate mobile device models to deliver cloud learning are explored with their benefits and challenges. Building on the mobile cloud learning platform, various mobility learning models are identified that enable instructors to adapt the curriculum, content and resources to individual learners. Several academic and corporate industry case study implementations referenced highlight how cloud learning has been successfully implemented for mobility. This reinforces the conclusion that learning is moving towards mobility, optimizing the learning experience of each student.

Keywords—Cloud computing, mobile computing, mobile learning

I. WHAT IS THE CLOUD?

Cloud computing is a term that has gained widespread use in today's culture, but what is "the cloud?" To put it in simple terms, the cloud consists of software programs, and applications that run on computers connected to the internet, rather than on your local device. It is a logical evolution of computing that was first imagined over fifty years ago by John McCarthy. After becoming frustrated with the process of using the mainframe computer at the Massachusetts Institute of Technology (MIT) in 1957, and subsequently joining MIT 1959, Mr. McCarthy wrote a memo to the Director of the Computer Center suggesting that teletypewriters be linked to the computer. This could allow several people to access the computer at the same time and lessen some of the frustrations stemming from the use of punch cards, and the long waits needed to see the results of computations submitted to the computer center. From this early concept of connecting multiple people to the same computer, he envisioned computing becoming organized similar to a public utility, where people pay for their usage [3].

It took decades for the computer, and communication technology to advance to a point where this concept has become a reality. We are currently in a transitional phase, as more and

more processing and storage of our personal data, and software we use are running on the internet, and not necessarily housed on our computing devices.

II. WHAT ARE CLOUD LEARNING AND MOBILE CLOUD LEARNING?

Since the cloud is really a network of data centers that are accessed through the internet, cloud learning can be looked at as learning content, and software that is housed in the cloud, and available through the internet. This content can be in the form of job aids for performance support, videos, live and recorded webinars, text documents, or any other ways that information can be made accessible. Entire courses can be accessed through the cloud, leading to the possibility of earning diplomas, certificates, certifications, and degrees completely through materials in the cloud. The key is that the cloud provides computing power, and data storage, lowering the requirements of access devices [2].

This leads us to mobile cloud learning that is generally differentiated by the type of device being used to access information. Cloud learning is thought of as being accessed by a desktop computer that is connected to the internet, while mobile cloud learning is viewed as learning that happens on a mobile device, whether, a phone, tablet, or laptop computer that is wirelessly connected to the internet. The latest definition focuses more on the learner's mobile experience, and less on the actual devices [5].

III. ADVANTAGES AND DISADVANTAGES

New concepts and new technology have both advantages and disadvantages for the institutions, and for the learners. Below is a table that summarizes some of the major advantages and disadvantages associated with mobile cloud learning.

Table 1. Advantages and Disadvantages

<u>Advantages</u>	<u>Disadvantages</u>
Can access content from any computer, or mobile device with an internet connection	Must have an internet connection to access
24-hour access—use whenever it is convenient	Different devices can have access problems to information, and smaller less

	capable devices may not display things as intended
Learning resources can be shared by many different providers	Data sent back and forth to the cloud can be intercepted by hackers, or data stored in data centers could become compromised
Since the cloud provides computer resources, simpler devices can be used to access	Risk of cloud service provider going out of business, which can lead to loss of resources, and loss of records
Can integrate social media as a tool to learning	The output a user receives can vary based on the browser they are using
Reduction of capital expenses in computer infrastructure and maintenance	Download and playback speeds can be negatively affected by the quality of the connection, and the processing speed of the device
Organizations can disseminate information quickly to all locations they service in a short time	Courses or Modules must be shorter as learners focus time is shortened
Available on a wide variety of devices	Design for mobile can add complexity to course and resource development

[4][5]

IV. CLOUD LEARNING SYSTEMS

The most popular cloud learning systems fall into three categories - Learning Management Systems (LMS), Student Management Systems (SMS), and Video Capture or Webcast Systems [7].

A. Learning Management Systems (LMS)

Wikipedia defines a Learning Management System (LMS) as “a software application for the administration, documentation, tracking, reporting and delivery of electronic educational technology (also called e-learning) education courses or training programs.”

LMS were initially developed as on-premise solutions. As the internet cloud started to gain popularity, several on-premise only LMS solutions evolved to offer a Software as a Service (SaaS) option. Other LMS providers entered the cloud learning market with solutions designed specifically with web-based mobile learning as the goal. Inquisiq R3 LMS by ICS Learning Group and WiZDOM LMS by G-Cube are offered as either on-premise or SaaS implementations. Joule from MoodleRooms (acquired by Blackboard) was developed to create virtual classrooms as an extension of the Moodle

platform. Examples of LMS solutions designed specifically as a cloud-based learning solution are Talent LMS, Litmos, and Latitude.

Many LMS systems have been more widely adopted in Corporate like Talent LMS and Litmos while others are designed for and more readily adopted in academic sector like Haiku and Joule LMS. Other LMS providers target niche markets or specific industry segments like Learning Evolution, a leader in the consumer and packaged goods (CPG) and retail industries. VTA Talent Management Suite by RISC is widely used in heavily regulated industries.

Other differentiating factors between LMS systems include their ability to offer online or offline study options, multi-language support, simplicity in use and administration, multi-device accessibility through PCs and mobile devices, cost, SCORM compliance and ability to integrate with other learning systems and tools.

If you want to get started with a free LMS option, consider OpenClass LMS from Pearson, a joint venture with Google or TalentLMS, an easy free plan for up to 5 users, 10 courses and up to 20 GB per download limit [8].

B. Student Management Systems

Edutech Wiki defines a Student management system (SMS) as “software to manage all day to day operations for a school.” Student Management Systems may also be referred to as Student Information Systems (SIS), Student Information Management Systems (SIMS) or Student records system (SRS). Functionalities between systems vary between providers. However many common features that are integral to the cloud learning experience of students include admissions and enrollment in classes, managing academic progress through grades and transcripts, tracking honors and standardized testing, tracking attendance and disciplinary actions, communications, and post-graduation alumni contact.

Just like the LMS systems, Student Management Systems can also be on-premise or in the cloud. A few examples of Student Management Systems available as web-based cloud deployments are Blackboard, Populi, Administrator’s Plus by Rediker, School Management Software by RenWeb, and School Minder by Hunter Systems.

C. Video Capture and Web Conference Systems

In cloud learning, video and web-conference plays a key role in instructional delivery. The video capture enables capturing of demos or webcasts for students to be able to view at a later time for review. Or, if a student’s schedule does not permit them to attend the live webcast, they can view a recorded webcast session at a more convenient time.

The web conference systems enable a virtual classroom to take place where all students and the instructor are connected through audio and a webcam. The instructor is able to share their screen, show presentations, share their desktop, and conduct demos. The instructor can also use a camera to show hands-on labs (such as a physics demonstration) to students attending virtually. The advent of video conferencing and video

capture has opened up opportunities for anytime anywhere learning. It has enabled instructors to deliver a class from halfway around the world! And it has brought in students from all over the world into one virtual classroom or educational experience.

Web-conference software examples offered in the cloud include more informal social media options such as Google Hangouts, Skype or Facetime to more robust tele-conference solutions used in Higher Education or Corporate such as GoToMeeting, WebEx, Adobe Connect or Microsoft Lync. Video capture software examples range from free downloadable options like HyperCam and Ezvid to more robust video and screen capture software such as Camtasia and Adobe Presenter.

V. DEVICE MODELS THAT EXTEND CLOUD LEARNING TO MOBILE CLOUD LEARNING

There are two main device models that enable mobile cloud learning - 1:1 where everyone uses the same device such as an iPad or "BYOD" otherwise known as bring your own device. A third model would be a hybrid of the two where some portions of a learning environment (such as a school) may adopt 1:1 for some classes and allow BYOD for other classes [7].

Variables that influence an institution's decision when evaluating which model to adopt include the number of students, socioeconomics, demographics, public versus private institutions and location (urban, suburban or rural). Public schools need to balance the goal of providing an equal educational experience to all students with the realities of taxpayers' willingness to support the initiative and the school's budget. In Higher Education, there is movement toward establishing a consistent learning environment with standardized tablets. At Florida's Lynn University all freshman are issued the same tablet while Illinois Institute of Technology provides a standardized tablet to all first year undergraduates [7].

Using mobile devices result in several learning benefits. Core curriculum delivery through mobile devices can create a more interactive learning environment and generate more engaging discussions on the topic. Students also get to use devices they enjoy and are comfortable using while reducing textbook costs up to 50% [7].

VI. MOBILE CLOUD LEARNING MODELS

Through the use of cloud learning, instructors are able to adapt the instructional approach to the individual student's learning needs. Cloud learning focuses on the curriculum, content and resources made available to the learner - providing interaction, critical thinking skills and complex problem solving that traditional textbooks and other resources lack [7]. Learning approaches that easily adapt to cloud learning for more effective instruction are assessment, remediation, test prep, indexes (or lists), guides, collaboration and supplemental materials.

Cloud-based curriculum and digital content enables teachers to create and adapt their curriculum and instructional delivery through a variety of learning models including [7]:

- Use of adaptive technology to personalize lessons to each student's knowledge and skills.
- Game-based learning that can be very effective in K-6 grades.
- Video lectures and lessons that students can replay sections of if they don't understand what was taught the first time.
- Interactive study aids such as digital flashcards and practice questions.
- Social media tools where students and instructors can interact by collaborating and sharing their ideas, knowledge and experiences.
- Digital textbooks which are interactive and engaging, immersing the student in the reading and learning experience.
- Interactive and adaptive test prep where the learning is more engaging and relevant

Supplemental materials can vary depending on the instructional outcomes and core curriculum already available. They can include examples, case studies, further reading and resources. Supplemental materials can also be used to deliver content in small "nuggets".

VII. APPLICATION AND USES OF MOBILE CLOUD LEARNING

A. Application and Uses in Education

Moodle is one Cloud Learning platform used extensively in the education sector around the world, yet still has resistance from both instructors and students. This resistance exists even with the move of Moodle to the cloud. However, Khalifa University has experienced significant benefits by moving their Moodle LMS from on-premise to Moodle Version 2.0 in the cloud. With the cloud implementation, Khalifa University enabled off-campus users to access the LMS through their mobile devices, such as smart mobile phones and iPads. As a result of the move, they were also able to shift their investment and resources from managing the technology to supporting the learners and teachers/professors. Since applications developed for Moodle 2.0 are the predominant paradigm for mobile development, Khalifa University was able to adapt and integrate their LMS with other systems to leverage mobile device functionality for communications and collaboration, cloud storage and Moodle access [4].

In the K-12 sector, mobile cloud learning deployments are typically phased in with pilot adoptions for either 1:1 or BYOD device models. Clark County School District, a large urban public school district has over 315,000 students and 357 schools. They began with 1:1 deployment to 12,200 students at nine Title I schools while simultaneously deploying BYOD to two of their schools. Since the 1:1 was too costly to deploy district-wide, they have now adopted a district-wide BYOD policy, giving schools the option to deploy BYOD and supporting them with technology infrastructure, learning integration and training [7].

B. Application and Uses in Industry

Mobile cloud learning can be a very effective way for businesses to provide training to their workforce, customers, and potential customers. The cloud facilitates easier deployment of training materials, especially to a geographically diverse population of employees, and allows those people to collaborate on their learning experiences [6]. This collaboration can lead to more functional teams, work efficiencies, and faster knowledge dissemination throughout the organization.

Cloud technology also allows the training personnel in a company to focus on training content, and deal less with hardware and software updates needed to deploy the training. They can also devote more of their budget to actual training, when they don't need to factor in purchases of software and equipment to deliver the training [4].

One example of using the cloud to help reach business goals is seen in SAP, a leader in software solutions for business. They have doubled in size to 67,000 employees in just 4.5 years, and needed a solution to their learning and development needs. They truly believe in talent development, evidenced by two of their corporate mantras: "Everyone is a teacher and everyone is a learner", and "Everyone is a talent". Both of these sayings point to the importance that management places on learning in their organization.

They also have a guiding leadership principle of ensuring customer success. They do this by developing their employees, leading to better products and services for their customers. To help their customers, they provide a cloud-based learning platform called Learning Hub, which hosts online courses in innovation.

SAP recently moved from a local LMS to a cloud-based LMS, and they used that opportunity to streamline the courses they offer, making sure what they did have provided a measurable business impact. This cloud-based LMS helps hold everyone in the company accountable for learning, and gives management metrics to help evaluate learning activity, and business results.

The cloud-based learning and development program helps them attain new talent, retain existing talent, and move their business forward by preparing them for the future [10].

Qualcomm is another example of a successful company using mobile learning to educate its workforce. They've created a suite of applications that can be accessed from a mobile device, and used during a new-hire's onboarding process. Rather than getting all the information at once, they can use the tools when they need them and at any time. Qualcomm also has a web platform called the "Qualcomm Journey" that employees can use to share their stories about work in video format, for others to see.

They too know that one way to increase the bottom line is through attraction and retention of great employees, and then giving them tools to increase their effectiveness once they are on board [11].

VIII. CONCLUSION

With the mass adoption of communication and collaboration through mobile devices in today's culture and younger generation, it is inevitable that learning leverages the power and community of mobile devices. However, organizations and academia still need to invest time and resource to make best use of cloud learning. It is important to invest in mobile learning that has a measurable impact. There is also a need to invest in research of best practices in applying pedagogy to mobile learning.

Keeping up with the rapid technology advancements of mobile devices and their capabilities can be challenging though. Wide availability and constant introduction of new apps can make standardization difficult. This is especially seen in the K-12 academic sector with technically savvy and creative instructors who like to try the latest app or tool available. For cost saving reasons, accessibility to free apps also drives app use decisions by educators. Meanwhile, the use of free apps or unapproved apps can generate concerns within the IT departments for security, confidentiality, maintenance and administrative reasons. Large enterprises and heavily regulated industries typically experience more restrictions when determining standard software, tools and apps that can be used across the company, slowing the adoption and use of leading edge mobile learning.

In both academia and industry, well-established curriculum have been built for the classroom that continues to be used. Although ultimately you can't completely replace existing training with mobile learning; they can be complementary, combining the best of both deliveries to optimize the learning experience for each student.

Acknowledgement

This paper is supported by the Oriental Scholar program of Shanghai Municipal Education Commission (TPKY052WMJ).

REFERENCES

- [1] N. Y. Asabere. (2012). Towards a Perspective of Information and Communication Technology in Education: Migrating From Electronic Learning to Mobile Learning. *International Journal of Information and Communication Technology Research*, 2(8), 646-649
- [2] M. Chen, Y. Ma, Y. Liu, F. Jia, Y. Ran & J. Wang. (2013). Mobile Learning System based on Cloud Computing. *Journal of Networks*, 8(11), 2572-2577. Doi, Available: <http://ojs.academypublisher.com/index.php/jnw/article/view/jnw081125722577>.
- [3] V. Rajaraman, Cloud Computing. *Resonance*, 19(3), pp. 242-258. doi: 10.1007/s12045-014-0030-1, 2014.
- [4] V. Ratten, Implementing Cloud Learning in an Organization: A Training Perspective. *Industrial and Commercial Training*, 44(6), pp. 334-336, 2012.
- [5] M. Wang, Y. Chen, and M. J. Khan, Mobile Cloud Learning for Higher Education: A Case Study of Moodle in the Cloud. *International Review of Research in Open & Distance Learning*, 15(2), p. 254, 2014.
- [6] J. Liao, M. Wang, W. Ran, and S. J. H. Yang. (2014). Collaborative cloud: a new model for e-learning. *Innovations in Education and Teaching International* 51(3), pp.338-351. Available: <http://dx.doi.org/10.1080/14703297.2013.791554>.
- [7] J. Halpin, C. Brown, Mobility & Cloud: Shifting Campus & Classroom to Cloud- and Mobility-enabled Learning Models. *Center for Digital Education Special Report*, 2013.

- [8] C. Pappas. (May 18, 2013). The Ultimate List of Cloud-based Learning Management Systems. *eLearning Industry*. Available: <http://elearningindustry.com/the-ultimate-list-of-cloud-based-learning-management-systems>.
- [9] Admin (Dec 8, 2010) 7 Learning Models for Mobile Learning *Mobl21*. Available: <http://www.mobl21.com/blog/08/7-learning-models-for-mobile-learning/>.
- [10] J. Dearborn, (2015). Learning at the Speed of Business: SAP Leads in the Cloud. *TD Magazine*. Available: <https://www.td.org/Publications/Magazines/TD/TD-Archive/2015/01/Learning-at-the-Speed-of-Business>.
- [11] R. Pyrellis, (2015). Chief Learning Officer Magazine - June 2014 - Qualcomm - Mobile-Friendly Learning (Jun 14).pdf. Available: [http://www.cedma-europe.org/newsletter%20articles/Clomedia/Qualcomm%20-%20Mobile-Friendly%20Learning%20\(Jun%2014\).pdf](http://www.cedma-europe.org/newsletter%20articles/Clomedia/Qualcomm%20-%20Mobile-Friendly%20Learning%20(Jun%2014).pdf).

Philip Robisch is the Product Technical Resource Manager at Hunter Industries, a global leader and manufacturer of irrigation, landscape lighting, and custom manufacturing products, doing business in 125 countries across the world. He is responsible for product training for all types of customers through the company's online training site. He developed a certificate program that allows customers to earn awards based on their level of participation in the training site, as well as in-person classes that focus on practical application of advanced irrigation control products. Address for correspondence: Philip Robisch, 1940 Diamond Street, San Marcos, CA 92078. Tel: 760-591-7146, Email: Philip.robisch@hunterindustries.com.

Becky Kirsininkas is the Registrar at Horizon University in San Diego responsible for student records management, course offering cycles, student enrollment and degree tracking. She is also responsible for project management and development of various academic projects including documentation of policies, procedures, handbooks, templates, job aides and guides to support accreditation. Becky has been the technical contact for administration, training and support of student management and LMS systems used in K-12 and Higher Education and taught High School Computer Applications. Her prior experience in the corporate sector includes technical curriculum development and education program management for the Microsoft Certified Trainer Channel and Microsoft Curriculum Strategies. Address for correspondence: Becky Kirsininkas, 3334 Avenida Hacienda, Escondido, CA 92029. Tel: 415-971-1871, Email: bkirsininkas@gmail.com.

Minjuan Wang is an oriental scholar at Shanghai International Studies University, China, a professor of Learning Design and Technology at San Diego State University, and a Program Manager for the Chancellor's office of California State University. Her research specialties focus on the sociocultural facets of online learning, and the design and development of mobile and intelligent learning. She has published peer-reviewed articles in *Educational Technology Research and Development*, *Computers and Education*, *Educational Media International*, *TechTrends*, and the *British Journal of Educational Technology*. She has also published book chapters on engaged learning in online problem solving, Cybergogy for interactive learning online, informal learning via the Internet, and effective learning in multicultural and multilingual classrooms. Address for correspondence: Dr MinjuanWang, 5500 Campanile Dr. PSFA 315, SDSU, San Diego, CA 92182-4561. Tel: 619-5943878 Email: mwang@mail.sdsu.edu.

Research on the Analytics Model Design of Online Learning Behavior

Jun Xiao, Minjuan Wang, Lamei Wang, and Bingqian Jiang

Abstract—Drawing on current research about online learning behavior, we constructed an analytics model, which consists of four major analytical stages: data collection, data organization, data analytics, and data application. The four stages form a continuous cycle during the implementation process. Meanwhile we identified several key factors of the analytics model for online learning behavior, including the design of data collection index, data collection and organization mode, as well as key techniques of data analytics. The analytics model of online learning behavior can support relevant systems, which are capable of collecting, analyzing, and extracting teaching data from mainstream learning platforms. Learning systems built on this model can also provide intelligent services for teachers and students in distant education.

Keywords— Knowledge and Data Technology, Adaptive and Learning Systems, analytics model, learning analytics system

I. INTRODUCTION

With the improvement of science and technology, online learning has significantly changed people's understanding towards education. In addition, the emergence of new technologies and ideas are propelling online learning forward. According to a survey report by the Sloan Consortium of the United States, many colleges and universities in the U.S. have made online learning an important part of their long-term development plan from 2010 onwards [1]. The trend is on a rise annually, and a majority of colleges and universities surveyed has realized the importance of online learning.

In 2012, elite schools such as Stanford, Harvard and the Massachusetts Institute of Technology took the same action of raising a trend of MOOCs, which drew an extensive attention

This paper is supported by “Shu Guang” award “MOOCs design and empirical research oriented Shanghai lifelong learning (13SG56)” from the Shanghai Municipal Education Commission and Shanghai Education Development Foundation. It is also supported by the 2014 Shanghai Education Scientific Research project “The Study of online learning mode for Shanghai lifelong learning (A1403)”. Besides, thanks for the support of Science and Technology Commission of Shanghai Municipality research project “Shanghai Engineering Research Centre of Open Distance Education (13DZ2252200)”.

Jun Xiao is with the Shanghai Engineering Research Centre of Open Distance Education, Shanghai Open University, Shanghai 200433 China. (phone: (+86)021 25653263; fax: (+86)021 25653263; email: xiaoj@shtvu.edu.cn).

Minjuan Wang is with the San Diego State University, San Diego, USA. (email: mwanj@mail.sdsu.edu).

Lamei Wang is with the Shanghai Engineering Research Centre of Open Distance Education, Shanghai Open University, Shanghai 200433 China. (email: wanglamei@shtvu.edu.cn).

Bingqian Jiang is with the Department of Education Information Technology, East China Normal University, Shanghai 200062 China. (email:51130104045@student.ecnu.edu.cn).

from global universities, enterprises, the whole society and even individuals. An increasing number of people started to register courses on mainstream MOOCs platforms such as Coursera, edX, Udacity, which have different focusing areas. With an expansion of the scale of online education learners, resources and interactions, more and more people choose online learning. At present, online learning has become one of the main learning modes [2]. Therefore, analyzing and evaluating learners' online learning behavior can help promote and improve learning outcomes. In addition, it can help to improve teaching practice and optimize the development of a lifelong learning platform. An essential part in developing an online learning platform is to meet the demands of big data processing.

II. LITERATURE REVIEW

In recent years, “learning analytics” has aroused a widespread interest in the education industry. As an education data analysis technique, learning analytics, based on the big data, has become an indispensable part of the development of online learning. The biggest benefits can be pursued through the discovery and understanding of the data's hidden information, as well as an efficient utilization of researches (for example, teaching intervention, learning prediction) [3],[4].

This literature review reveals that current studies on online learning behavior analytics has mainly focused on the following aspects:

A. Description and feature analysis of online learning behavior

Analysis of online learning behavior is an important pre-requisite for carrying out the design of online learning system and the development of online education resources. In order to improve the efficiency of online learning, to motivate learners, and to improve learners' interests, it is necessary to have a clear understanding of the basic concept of online learning behavior, and the characteristics and styles of online learning behavior among individuals and groups [5][6]. In current research findings, quite a few researchers have carried out investigation on online learning behavior among specific groups, and they also focused on the patterns, characteristics and frequencies of learners using online self-directed learning. Moreover, these studies analyzed the status quo and level of online learners and explored causes [6].

B. Research on the data model of online learning behavior

Experts in different disciplines have different opinions

towards the models of analyzing online learning behavior. For example, Siemens (2010) believes that learning analytics is composed of several stages, including data collection, data analysis, data prediction and data adjustment. He constructed a linear analytics model that includes these four elements. According to Elias (2011), learning analytics has three stages, namely data collection, information process and knowledge application, which form a continuous cycle. Learning analytics also has six activities, which are acquirement, selection, gathering, prediction, use, and optimization. Based on these components, she proposed an application model of learning analytics that is relatively more complicated. Greller and his colleagues (2012) built a theoretical model of learning behavior analysis in terms of data sources, analytical modes, constraint condition, competitiveness and interested parties. In recent years, researchers have started to apply the learning analytics techniques to the online learning analysis of MOOCs.

Summarizing theories and models from existing studies, learning analytics has core elements such as targets, objects, restraints, data resources, and processing methods, which reflect the internal and external conditions for carrying out learning analytics. The process of learning analytics consists of three stages: data collection, data processing, and data application and feedback, which form a continuous cycle during the implementation process [7][8].

C. Research on data collection techniques and system implementation of online learning behavior

Online learners' behavior data are acquired from Web blogs, network sniffing, questionnaires, platform database, and web data mining technique, mobile Agent intelligent agent technique, standard SCORM online learning technique, and electronic portfolio technique. For example, Hummel adopted a way of obtaining relevant data of learners' learning behavior from the database access records and server access logs of online learning systems. Wu and his colleagues from Xi'an Jiaotong University created an algorithm using resolution function for attribute reduction, based on the rough set theory. They discovered that learners' key feature dimension is merely 1/4 of its original dimension. Besides, learners' key feature can be detected automatically, and the objective relation between personality characteristics and learning strategy can also be revealed. This algorithm can simplify the data analysis [9].

D. Research on the analysis techniques of online learning behavior

Online learning behavior analytics is a newly-developing educational research, which adopts techniques including business Intelligence, network analysis, education data mining, academic analysis, etc. These intelligent techniques can be used to analyze and process massive data. Among which, data mining method has been applied in the online learning platform, which initiates the research on learners' behavior using analytics technique. Apart from the above common analysis techniques, the analysis technique of online learning behavior has been taking in and integrating other techniques and methods, including analytic methods of social network,

discourse, and content. Introduction of these new analytic methods has largely enriched the data processing approaches and strategies for analyzing online learning behavior [10][11].

E. Evaluation research of online learning behavior

Researchers have also paid close attention to the evaluation research of online learning behavior. For example, Xu and his colleagues (2003) have proposed an evaluation technique for the interaction of online education and learning environment. He believes that interaction is the core of online learning, while current online learning lacks feedback. Websites using the analysis method that has been applied to business websites, together with an analytics model, using mining technique, serve as the final evaluation methodology. Outtaj and his colleagues (2007) divided online learners into different types, and applied different evaluation standards to analyze the different types of online learners [12]. Although this research has touched on the fact that various aspects should be considered when evaluating a student, it only studied the interactions between learners, such as interactions among learners, interactions between learners and the learning platform.

In conclusion of above research results, there are still some shortcomings in the analysis research of learning behavior [13]:

Firstly, an absence of the comprehensive collection and analytics of a multi-platform, multi-terminal data, as well as data of students' dominant behavior. Currently, research on the data collection of online learning behavior is mainly dependent on a fixed learning platform, and data obtained tend to have an "apparent" mode, which lacks data collected from present multi-platform and multi-terminal learning behavior, as well as that of learners' dominant behavior and physical signs. Therefore, it is relatively difficult to analyze and share via different platforms.

Second, the index of learning behavior analytics and evaluation index needs alignment. Online learning behavior analytics is a comprehensive, dynamic and systemic process. Although there is a relatively in-depth research, in theory, on the existing index system of online learning behavior, these theories cannot properly serve as the evaluation objectives of specific online learning behavior analysis. The key problem lies in the fact that the index cannot be properly matched to the evaluation objectives.

Third, there is a lack of smart feedback from the evaluation results of online learning behavior. Although a majority of current online learning platforms provide evaluation functions, such as online homework and testing, the evaluation results are not ideal. Most of the information provided by existing researches is periodical or conclusive, which lacks the individualized learning resources for learners, or recommendations for their study.

Fourth, a lack of in-depth analysis for big data. Due to the fact that not a technique is capable of extracting learners' learning behavior from a virtual environment, most of the current research discusses online learning only by observing and analyzing the backend database of a few web-based teaching platforms. And current researches on online learning behavior appear to be a simple statistical description of data,

which lack the in-depth mining of big data.

III. MODEL DESIGN OF ONLINE LEARNING BEHAVIOR ANALYSIS

Acting upon the aforementioned gaps, integrating the characteristics, method, and technique of online learners' behavior analysis, and the objective function of learning analytics, we propose an online learning behavior analysis model as shown in the Fig.1. The process of behavior analysis will occur in four phrases: data collection, data organization, data analysis, and data application, which form a continuous cycle during the implementation process [14].

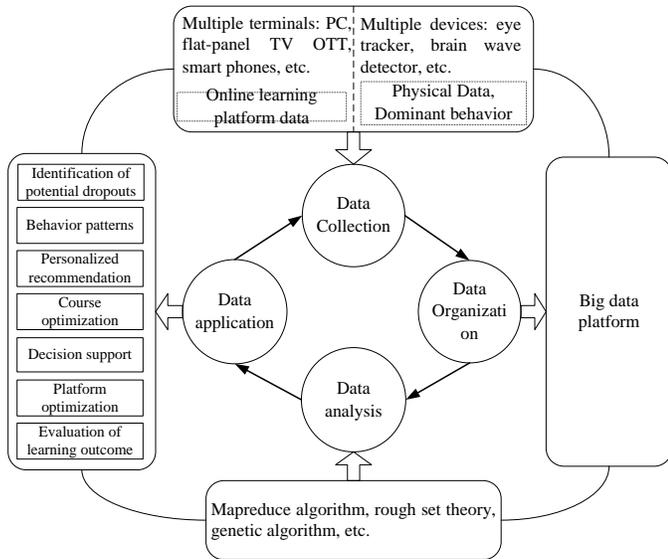


Fig. 1 Analysis model of online learning behavior

(1) Data collection stage: In this stage, based on the online learning feature of “teachers and students separating from each other”, a wide selection of data resources are included, which are data from different learning terminals (PC, tablet, TV OTT, smart phones, etc.) and a third party data platform. Techniques used are not only the common ways of obtaining virtual data, namely Web blogs, database, etc., instruments like eye tracker and brain wave detector are also adopted to get data of learners’ physical sign, all of which contribute to a comprehensive and objective data. This has set up a basis for the following in-depth analysis.

(2) Data organization stage: It is unavoidable to have repetitive or invalid data among those obtained from collection stage, therefore the sorting and cleaning process is needed, and the final data will lead into the developed big data platform.

(3) Data analysis stage: In this stage, certain algorithm and analysis model will be adopted to process the data from Big Data. And supports will be provided for all kinds of decisions and services needed in the next stage.

(4) Data application stage: In this stage, a wide range of services and supports are provided, by means of the operation results from data analysis stage, including identifying potential dropouts, establishing learners’ behavior patterns, providing personalized recommendation service for students, offering decision support for education administrators, offering support to the optimization of both platforms and resources, carrying

out effectiveness evaluation for learners etc.

IV. KEY FACTORS IN THE MODEL DESIGN OF ONLINE LEARNING BEHAVIOR ANALYSIS AND SOLUTIONS

Several key factors need to be considered in the model design of online learning behavior analysis, including data collection index design, modes of collecting and organizing data, as well as key techniques of data analysis. The analytic system of the above online learning behavior analysis model is able to collect various kinds of learners’ behavioral data in the digital learning process, and carry out a rapid analysis to all learning behavior data, such as learners’ browsing habit, ways to open webs, in the learning process [15]. The system can also have a direct access to a range of third party learning platforms, and analyze learning behavior on the basis of historic data.

A. Data collection index design of online learning behavior

In the design of analysis index, we adopted the concept of Objectives and Key Results as the guiding ideology for the index design, i.e. applying the cycle of “target decomposition – quantification of key results – evaluation of measures” in the index system design. It has overcome the limitation caused by a low compatibility between the evaluation objective and the former index design method of using Key Performance Indication (KPI) as the core. Therefore, it is a more reasonable and scientific index design.

Based on the current researches on online learning behavior index, together with specific evaluation objectives of online learning behavior analytics, we made the online learning behavior data collection model [16]. This model caters to teachers and students with seven evaluation objectives, which is shown in Fig. 2.

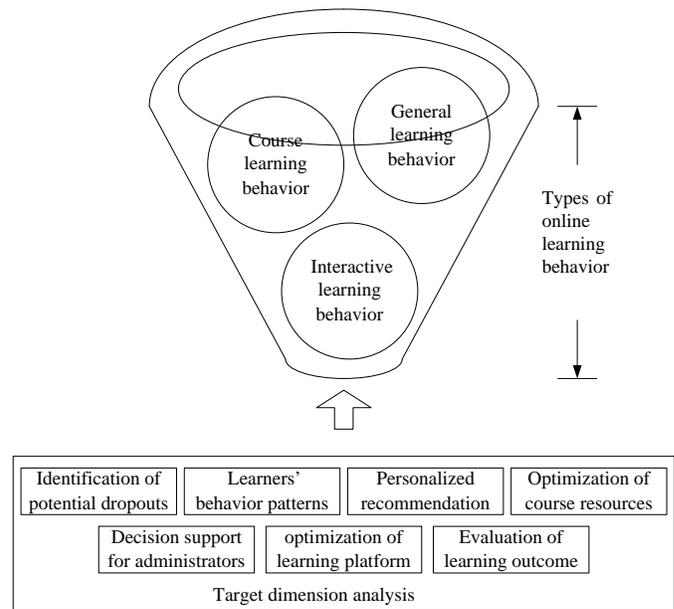


Fig. 2 Data collection model of online learning behavior

II -level evaluation indexes includes the above seven objective dimensions: identification of potential dropouts, learners’ behavior patterns, personalized recommendation, evaluation of course resources, decision support for

administrators, optimization of learning platform, and evaluation of learning outcome. There may be intersections, for example indexes that reflect learners' behavior pattern, learning style and learning path can also be used as specific indexes that support personalized recommendation service. Therefore, according to the basic principles of activity theory and the principle of behavioral science management, we have applied three dimensions, learners' general behavioral habit, interactive learning behavior and course learning behavior, to collect online learners' behavior data, on the basis of the above seven objective dimensions. Moreover, after the comprehensive analysis of current evaluation indexes, classical index group is formed, by means of a preliminary selection of those frequently used indexes, adopting statistical method for frequency. On the basis of that, the most representative and non-repetitive indexes have been selected by applying clustering analysis method in order to analyze the chosen evaluation indexes [17]. Furthermore, adjusting and revising index sets for several rounds, and formulating a detailed second-level index, as shown in Table 1.

Table 1 Data collection index of learners' behavior

I-level Index	II-level Index(encoding)	Meaning
General Behavioral habit	Learning Path(1)	Learning order
	Learning Properties(2)	Length, time and equipment of learning
	Ways of Collecting Information(3)	
	Ways of Processing Information(4)	Quantity and quality of making notes
	Ways of Distributing Information(5)	Number of posts and learning logs, etc.
	Concentration Level in Learning(6)	Degree of concentration
Interactive Learning Behavior	Involvement in online FAQ (7)	Effective times for questioning or answering
	Participation in BBS discussion (8)	Effective times for questioning or answering
	Usual interactive ways(9)	Synchronous or asynchronous
	Usage rate of different interactive tools(10)	
Course Learning Behavior	Learning progress of the course(11)	Learning progress of individuals and groups
	Hit rate of different types of courses (12)	Hit rate of textual course, audio course, video course, image course, etc.
	Hit rate of different theme courses (13)	Teaching contents, course tests, course assignments, etc.
	Learning outcome(14)	Study on the academic achievements, etc.

Finally, encoding these II-level Indexes, and making an independent assortment of them to have correspondence with the above seven evaluation objectives (mode sets), as shown in Table 2.

Table 2 Analysis index catering to evaluation objectives

Analysis Objectives	Index
Identification of potential dropouts	1、4、5、6、7、8、11
Learners' behavior patterns	1、2、3、4、5、9、10
Personalized recommendation	2、3、12、13、14
Optimization of course resources	7、8、11、12、13、14
Decision support for administrators	3、4、5、6、9、10、12、13
Optimization of learning platform	1、5、6、9、10、12、13
Evaluation of learning outcome	2、4、5、7、8、11、14

B. Data collection method of online learning behavior

Analytics data of learners' behavioral process include learners' personal information, learning time records, distribution of learning areas, academic achievements recording, and records about other objective learning behavior. These data can be collected, quantified and stored in the behavioral database. It can be checked up in many ways, such as from the database backup of production system, or from the real-time/ semi-real-time recording logs, or on a regular or periodical basis, or from an event-driven way, or from an initiative order at the frontend, or pushing data automatically to the analysis page. As for the data obtained, system can establish a correlation between them, so as to carry out an in-depth correlation analysis, and provide a visual multidimensional display at the frontend by means of the report chart. In the meantime, learners' various learning behavior using "TV OTT, PC, tablets and smart phones" are connected, which will help collect data of individual user's learning behavior in different periods, as well as that of learning modes.

In light of learners' diversified learning behavior, various methods for leading third party learning system data should be provided as well, for instance database, log files, excel, web service interface, etc., in order to support a real-time process of big data. Therefore, we need to set up an a statistical analysis module for inquiring and invoking, based on the web service interface modes, by means of database backup, data importing function of the administrative page. The system imports the third party Excel data via the frontend administrative page, and shows the results after analysis. Users can have a direct access to the third party SQL database, or isomorphic Mongo database, they can also have a direct access to the third party data source through Web Service interface.

In data collection, considering the traditional educational measurement and ways of data collection from online learning platform, on the one hand, learning process data from learning platform is collected, while on the other, physical signs data of the students' dominant behavior should also be collected by intellectual devices, then using different algorithms to manage and analyze different types of data, which will reflect the students' factual thinking ability and learning.

The procedure of behavioral data collection and analysis in the learning process is shown in Fig. 3.

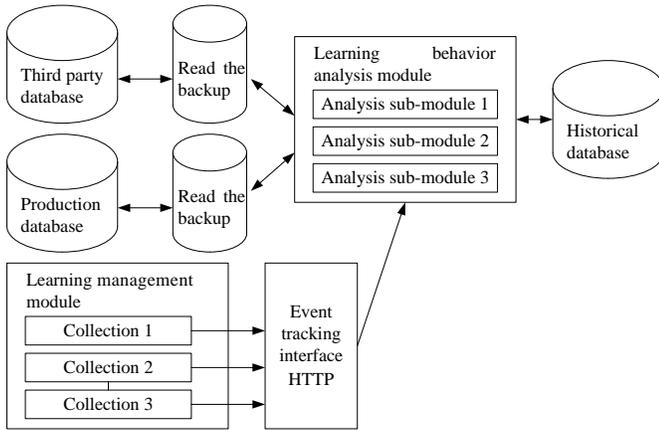


Fig. 3 The procedure of behavioral data collection and analysis in the learning process

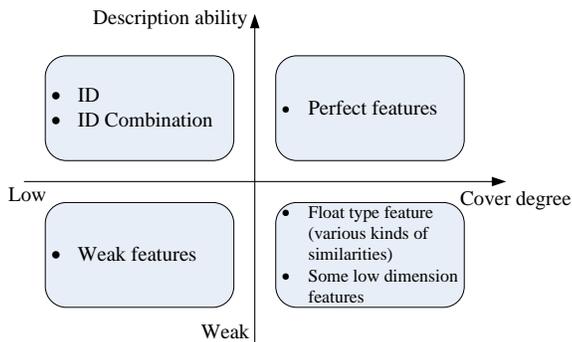
C. Key techniques and algorithms for analyzing online learning behavior data

The first step of online learning behavior analysis is to collect and sort out learning behavior data, and at the same time carry out an in-depth data mining and behavioral analysis using all kinds of algorithms, which mainly include feature sets, Mapreduce algorithm and theme mining algorithm[18][19], etc.

1) Feature data sets

Feature data sets contain the static data and application data of the learning resources, learning state data and so forth, which respectively correspond to data sources like historical data, real time data, attribute data, etc. Feature data sets need to go through several processes, namely pre-processing, feature extract and feature selection, as well as the final machine learning, in order to get the learning analytics model.

The core of feature set algorithm lies in the setting of the feature sets, as Fig.4 shows. Based on the cover degree of the feature sets (X axis) and description ability, learning analytics system trains data from each quadrant adopting different algorithms so as to obtain a model with high degree of fitting[20].



First quadrant: less general amount; a lot of Bias features
 Second and third quadrant: a relatively large amount; complement each other
 Fourth quadrant: weak influence

Fig. 4 Setting of the feature sets

2) Mapreduce Algorithm

Mapreduce technique is a typical example of non-relational data management and analysis techniques, which include three

levels of contents: distributed file system (DFS), parallel programming model, and parallel execution engine. Mapreduce technique is a concise parallel computational model. It can solve problems like expansibility and fault tolerance in the system level, and can also carry out a parallel execution automatically in the elastic large-scale cluster by accepting the Map function and Reduce function written by users. Therefore, it can process and analyze the mass education data.

3) Theme mining algorithm

Theme mining algorithm is a technique widely used in learning analytics, which is applied in both context targeting and behavior targeting. In the analytic system of learning behavior, we choose the supervised theme mining algorithm to map the page content onto the label system that has been defined in advance, rather than in an unsupervised way. The frame is to get to know learners' learning interests according to the learners' historical visiting behavior, and set up an analytics model. Behavior analysis is important because it provides a general way in making the most out of the online users' logs to have an in-depth analysis of teaching. Therefore, the frame, algorithm and evaluation index of behavior targeting establish the substantive characteristics of online data driven analysis. If you regard context targeting as the one that aligns with the user's single access behavior, behavior targeting can be regarded as the integration of a series of context targeting. Therefore, context targeting lays the foundation for behavior targeting analysis, and each type of context targeting can have its correspondent behavior targeting mode.

V. CONCLUSION

The online learning behavior system we described in this paper can collect data from different approaches, including learning process data collection from online learning platform, and sort out typical physical signs data from the actual scenario to conduct a comprehensive calculation and analysis. The system can also analyze data from multiple terminals and across different screens, which fills up the gap of cross-screen learning behavior analytics in the education industry. It aims at providing perceptible learning resources and recommendation services for learners, and also to assist with personalized learning. In the near future, we will analyze some learners' behavior data in the Lifelong Learning Platform for Shanghai Citizens to verify the feasibility of this study, and find out problems and rules existed in the current online learning situation. Besides, we will also make adjustments and improvements to boost the learning outcomes, and at the same time improve system function and analytic model according to the application situation.

ACKNOWLEDGMENT

This paper is supported by "Shu Guang" award "MOOCs design and empirical research oriented Shanghai lifelong learning (13SG56)" and the Oriental Scholar program (TPKY052WMJ) from the Shanghai Municipal Education Commission and Shanghai Education Development Foundation. It is also supported by the 2014 Shanghai

education scientific research key project “The Study of online learning mode for Shanghai lifelong learning (A1403)”. Besides, thanks for the support of Science and Technology Commission of Shanghai Municipality research project “Shanghai Engineering Research Centre of Open Distance Education (13DZ2252200)”.

REFERENCES

- [1] Hong Yan, Tang Hui, Liang Linmei. “The New Trend of American Higher Online Education-2010 & 2011 Sloan Consortium Survey Overview,” *Distance Education in China*, pp. 40-45, Jan. 2013.
- [2] China Internet Network Information Center. The 33rd China Internet network development state statistic report [EB/OL].2014:<http://www.cnnic.net.cn>.
- [3] Gu Xiaoqing, Zhang Jinliang, Cai Huiying. “Learning Analytics: The emerging data technology,” *Journal of Distance Education*, no. 208, pp.18-24, 2012.
- [4] Zhu Zhiting, Shen Demei. “Learning Analytics: The Energy of Smart Education,” *E-Education Research*, no.5, pp. 5-12+19, 2013.
- [5] Li Fengqing, Yang Shulin. “Higher Education in the Information Age: Future Trends and Challenges-The horizon report of NMC,” *Modern Distance Education*, no. 5, pp. 38-42, 2011.
- [6] Wei Shunping. “An Analysis of Online Learning Behavior and Its Influencing Factors,” *Open Education Research*, vol. 18, no.4, pp. 81-91, 2012.
- [7] Peng Wenhui, Yang Zongkai, Tu Qingshan.” The Survey and Analysis of Online Learner’s Learning Behavior,” *China Educational Technology*, no.12, pp. 52-56, 2007.
- [8] Pen Wenhui, Yang Zongkai, Huang Kebin.” Analysis and Model Research of Online Learning Behavior,” *China Educational Technology*, no.10, pp. 31-35, 2006.
- [9] Yang Jinlai, Zhang Yixiang, Ding Rongtao.”Online Learning Behavior Monitoring Based on Online Learning Platforms,” *Computer Education*, no.11, pp. 65-68, 2008.
- [10] Yang Jinlai, Hong Weilin, Zhang Yixiang.”Research and Practice of Real-time Monitoring of Network Learning Behavior,” *Open Education Research*, vol. 14,no.4, pp. 87-92,2008.
- [11] Hu Yunan, Lv Zhihui.”Design and Implementation of Learning Behavior Collection Based on SCORM,” *Computer Engineering and Applications*, no.22, pp.106-108, 2004.
- [12] Gao Yi, Shen Ruimin.”Learning Behavior Analyzing Center Based on Open E-Learning Platform,” *Computer Engineering*, vol.30, no. 15, pp. 86-88, 2004.
- [13] Jia-JiunnLo,pai-Chuanshu. Identification of learning styles online by observing learners browsing behavior through a neural network[DB/OL].32ndASEE/EIIEEFrontiersinEducationConferenece
- [14] George Siemens.What are Learning Analytics [DB/OL]. <http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics/>,2011-01-16.
- [15] Elias, T. Learning Analytics: Definitions, Processes and Potential[EB/OL].<http://learninganalytics.net/LearningAnalyticsDefinitonxProcessesPotential.pdf>, 2011-01-16.
- [16] [Wolfgang Greller. Learning Analytics framework [EB/OL]. <http://www.greller.eu/wordpress/?p=1467>, 2012-05-12.
- [17] Yassine Tabaa,Abdellatif Medouri. Karin Anna Hummel, Helmut Hlavacs. Anytime, Anywhere Learning Behavior Using a Web-Based Platform for a University Lecture[DB/OL]. www.ani.univie.ac.at/~hlavaes/Publications/ssgrr_winter03.Pdf.
- [18] Xu Lei, Claus Pahl. An evaluation technique for content interaction in Web-based teaching and learning environments [DB/OL].In Proceedings of The 3rd IEEE International Conference on Advanced Learning Technologies.<http://citeseerx.ist.Psu.edu/viewdoc/summary?doi=10.1.1.100.5701>
- [19] Wu Xi Yuan, Zheng Qinghua.” An Attribute Reduction Algorithm to Find Learner’s Key Characteristics Based on the Discernible Function,” *Journal of Xi’an Jiaotong University*, vol. 42, no. 12, pp.1455-1458, 2008.
- [20] Benaceur Outtaj, Raehida Ajhoun. Towards a model for evaluating the e-learner’s behavior [DB/OL].ICTA’07,http://www.esstt.mu.tn/utic/tica2007/sys_fjles/media/s/docs/P27.pdf

Jun Xiao is an associate professor of Shanghai Engineering Research Center of Open Distance Education, Shanghai Open University, China, the visiting scholar in department of computer science, San Diego State University, USA, and also the committee member of China Elearning Technology Standardization Committee. His research specialties focus on learning analytics, lifelong learning, digital learning system and educational resource repository research. He has led many large-scale Research and Development projects, such as Shanghai Educational Resource Center, Shanghai Lifelong Learning Network, Shanghai Learning Network, and this Cloud-Based Intelligent Learning System. He has published more than 30 articles on major publications, and he is the author of 5 books. Address for correspondence: Dr Jun Xiao, Learning Square, Room 505, No.288 Guoshun Rd, Shanghai, China. Tel: (+86) 021-25653263 Email: ecunxj2003@163.com

Minjuan Wang is an oriental scholar at Shanghai International Studies University, China, a professor of Learning Design and Technology at San Diego State University, and a Program Manager for the Chancellor’s office of California State University. Her research specialties focus on the sociocultural facets of online learning, and the design and development of mobile and intelligent learning. She has published peer-reviewed articles in *Educational Technology Research and Development*, *Computers and Education*, *Educational Media International*, *TechTrends*, and the *British Journal of Educational Technology*. She has also published book chapters on engaged learning in online problem solving, Cybergogy for interactive learning online, informal learning via the Internet, and effective learning in multicultural and multilingual classrooms. Address for correspondence: Dr MinjuanWang, 5500 Campanile Dr. PSFA 315, SDSU, San Diego, CA 92182-4561. Tel: 619-5943878 Email:mwang@mail.sdsu.edu.

Lamei Wang is a researcher in Shanghai Engineering Research Center of Open Distance Education at Shanghai Open University, China. Her research specialties focus on learning behavior analysis, and the design of E-learning system. She has published articles in China Educational Technology, Chinese Education Informationization. Address for correspondence: LameiWang, No 288, Guo Shun Road , Shanghai, China. Tel: (+86) 021-25653454 Email:wanglamei@shtvu.edu.cn

Bingqian Jiang received the B.S. degrees in educational technology from East China Normal University. She is currently working towards the M.S. degree in the Department of Education Information Technology, East China Normal University. Her research interests include learning technologies, learning analytics, and lifelong learning. Address for correspondence: Ms. Bingqian Jiang, Room 405 Computer Building, ECNU, No. 3663 North East Zhongshan Rd., Shanghai, China. Tel: (+86) 021 25653454 Email: jbq6888@163.com

Student anxiety awareness through a bio-feedback device as a significant support to educational activities

Hippokratis Apostolidis¹, Thrasyvoulos Tsiatsos¹, Minjuan Wang²

¹ Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
Tel.: +30 2310 998990, E-mail: {aposti, tsiatsos}@csd.auth.gr

² Learning Design and Technology, San Diego State University; Shanghai International Studies University
Tel.: +16195943878, Email: mawang@mail.sdsu.edu

Abstract—This paper describes an initiative in embedding students' affective feedback (i.e., frustration, boredom, anxiety and others) into learning tools and into emotional-appraisal learning models. The physiological reactions of the body such as conductivity of the skin, skin temperature, heart rate and respiration rate, produce bio-signals that can be classified into basic human emotional states. This work is the first approach in the direction of combining the cognitive models with the adaptive pedagogical patterns into a supportive reappraisal model. Anxiety state recognition and regulation may be approved an important factor in achieving better performance. This article describes our broad approach towards creating a modern and supportive learning environment.

Keywords— Affective computing, Mobile learning, Bio-signals, Bio-feedback, Emotional regulation, Emotional intelligence

I. INTRODUCTION

The emerging research on student emotions in classrooms focuses on wide areas of affective response. In their well-cited study, *Affective e-Learning: Using "Emotional" Data to Improve Learning in Pervasive Learning Environment*, Shen, Wang, and Shen [1] used emotion detection technologies from biophysical signals to explore how emotion evolves during learning process and how emotion feedback could be used to improve learning experiences. They described a cutting-edge pervasive e-Learning platform used in a Shanghai online college and proposed an affective e-Learning model, which combined learners' emotions with the Shanghai e-Learning platform. The "affective states or student emotions" in their paper are called cognitive-affective mixtures in this paper [2].

Various models have been proposed to explain the negative effects of stress. These models assume that stress includes activation of considerations unrelated to the specific learning task limiting students' engagement to the learning activity and thus causing greater effort. Students who are worried about failure cannot focus on their learning process.

Pekrun has classified academic emotions into categories that include achievement emotions, social emotions, and epistemic emotions ([2], [3]). Achievement emotions (e.g., contentment, anxiety, and frustration) are linked to learning activities (e.g., homework, taking a test) and have a result of

success, failure, etc. Social emotions such as pride, shame, and jealousy reflect the fact that educational activities are socially situated. Finally, epistemic emotions arise from cognitive information processing, such as surprise when novelty is encountered or confusion when the student experiences inconvenience. According to the control-value theory, these academic emotions arise from cognitive appraisals of control over the learning task and value in the learning activity, with reciprocal connections between the emotions, their antecedents, and their consequents ([2], [4]). Other research has focused on a more in-depth analysis of a smaller set of emotions that arise during deep learning in more restricted contexts and over shorter time spans, from 30-min to 1.5 h ([2], [5], [6], [7], [8]; [9]).

In general it could be considered that affect recognition can significantly improve a tutor's instructional design strategy e.g., when the tutor realizes that a student is bored he/she tries to create anxiety in the sense of engagement to the current activity. Observing learner's affective state continuously, a skilled mentor (experienced teacher or an intelligent tutor agent) may use that awareness, along with knowledge about cognitive progress, to reason about a series of student actions and interventions, not simply a single-shot action or interaction, but as an ongoing and evolving relationship. It is very important for the student to recognize when s/he is overly stressed and uses techniques to mitigate it so that it does not become overwhelming. One of the techniques used to mitigate that stress is the "Diaphragmatic Breath". This technique is well known in psychology, and it can be applied to all people and under any circumstance.

This paper presents a bio-feedback device which classifies human anxiety, by evaluating bio-signals from skin conductance, skin temperature, heart rate and respiration rate. The main motivations for this research are (a) a broad approach of the application of this device as a supportive tool of various educational activities and (b) to present further and more detailed possible use case scenarios where anxiety awareness through the bio-feedback device could be applied.

In this paper we begin by describing related research works. Then we present the affective computing methodologies used by this device. In the following section we present the bio-feedback device and how it works. We then describe

educational activities where the bio-feedback device was used and that brought to light the need for anxiety awareness. Afterwards we explore areas for further use of the device by presenting usage scenarios. Plans for future work are included in the conclusions.

II. RELATED WORK

There are many studies about methods of emotion appraisal, as mentioned in the bibliography. Emotion inquiry is still a scientific field where there is not a commonly accepted methodology. Many researchers suggested that the affective computing methods could be separated into two main categories a) observed methods and b) physiological methods [10].

Observed methods focus on facial and vocal emotion recognition. Also they try to detect emotions from gestures. Physiological methods could be considered as the bio-signal classification. The human bio-signals are produced by the physiological reactions of the human body as a response to world environment stimuli [10]. The AutoTutor project [11] applies observation methodology affective states detection through facial expressions, dialogue patterns, speech intonation and body movements. Another system applied real-time facial expression analysis to inform the design of the video recommender system and suggest meaningful recommendations of unseen videos ([10], [12]). Automatic facial expression analysis was also applied in a study of Google search [13]. The author found that during the search, surprise was the most frequently expressed emotion, followed by neutral, sad, fear and happy [10]. However, it is suggested that the facial emotion recognition is influenced from the context in which human emotions are occurring. Therefore, some researchers suggest that an analysis of human facial expressions could be more accurate if it would be combined with a synchronous context interpretation ([10], [14], [15], [16]).

Another observed method of affective computing is the analysis of verbal communication. In emotion stimuli conditions, speech is then characterised by loudness, increased speech rate and strong, high frequency energy ([10], [17]). In a research work, a set of affective features was extracted from multimedia audio content and was annotated using a set of labels with predetermined affective semantics. The audio features that consisted of speech, music, special effects and silence, were analysed in terms of the affective dimensions of arousal and valence ([10], [18]). Similarly, in another work, video content was modelled using a selection of low level audio (signal energy, speech rate, inflection, rhythm duration, voice quality) and visual features (motion) ([10], [19]).

An ambiguous field of observed affective computing methods is body movements and gestures. Some studies suggest that these are indicative of the intensity of emotion, but not its type, but others are assigning certain body movements with specific emotions in a reliable way ([10]; [20]; [21], [22]). Body movements, and specifically hand

gestures ([23], [24], [25]), have recently attracted the attention of the HCI community ([10], [26]).

In physiological methodology of affective computing the emotion recognition is detected by collecting bio-signals from brain activity, heart rate, skin conductance, respiration rate, skin temperature, and other physiological reactions. An example of a neuro-physiological instrument for detecting emotions is a LifeShirt sensor system that can be used for monitoring cardiovascular, respiratory, metabolic and other physiological effects of physical or emotional stress [27]. This is an example of wearable sensory system. The wearable sensors are a trend in physiological sensory systems in order to accomplish the request for unobtrusive sensors. In another research the human heart rate is measured with a touchless method using a mobile phone camera [37]. In the same direction there is another research work named "non-contact automated cardiac measurements using video imaging and blind source separation [38]. These projects could be considered as a combination between observed and physiological methodology.

III. AFFECTIVE COMPUTING METHODOLOGIES USED

Our main effort was to use as less obtrusive sensors as possible without decreasing the sensor sensitivity and accuracy. The presented bio-feedback device is using four affective computing methods skin conductance, skin temperature, heart rate and respiration rate.

A. Skin Conductance (SC or GSR)

This technique is associated with the change of electrical properties of the skin when external voltage is applied. This is a test of the sweat function, which measures the change in the conductivity of the skin during the flow of low voltage current after a stimulus. The recording of the conductivity (or the inverse of conductivity i.e. resistance) is based on the application of external constant voltage to the skin [28]. The edges of the human body (hands and feet) have a very high proportion of sensory nerves endings and so they become ideals for the application of skin resistance measurements.

B. Photoplethysmography (PPG)

Photoplethysmography (PPG) is a simple and low-cost optical technique that can be used to detect blood volume changes (BVP) in the microvascular tissue [29]. The PPG technology [30] requires a light source to illuminate the skin and a photodetector to measure the small variations of light reflected due to blood volume variations. These small variations create a normal, pulse waveform in every cardiac pulse. In every heartbeat, more blood is driven to the skin, which reflects the red light and absorbs other colours, so the reflection of light is greater. Heart rate (HR) is measured by the small variations in light intensity and it is an important derivative of the PPG method.

C. Skin Temperature

Various cognitive and physiological functions of the human body can affect the skin temperature. In particular

stressful situations may cause a reduction of skin temperature. This reaction is known in the literature as a consequence of a global phenomenon called «Fight or Flight».

D. Respiration Rate

Respiration rate is the number of breaths a person takes per minute. The rate involves counting the number of breaths for one minute by counting how many times the chest rises. Usually stress leads to rapid breath, which increases the respiration rate.

IV. DEVICE DESCRIPTION

The bio-feedback device (Fig. 1) is designed and developed in the Multimedia Lab of our Computer Science Department. So far it is used in demanding learning activities such as student oral or written examinations.

This device utilizes four physiological techniques including Skin Conductance (SC) or Galvanic Skin response (GSR), Skin Temperature (TEMP), Heart Rate (HR) and Respiration Rate (RR) measurement in order to collect human bio-signals. These bio-signals can be classified to human emotions. This device is based on an open source electronics prototyping platform Arduino duemillanove (<http://www.arduino.cc>) which is used as an analogue to digital converter. So far the bio-feedback device is connected to a personal computer via a USB cable or wirelessly (Bluetooth). A new mobile version is under development which will utilize the device connection to mobile Android smart phones or to Android tablets.



Fig 1. The bio-feedback device

The computer application receives bio-signal measurement values and stores them in an online database for recordkeeping. The online values are classified using machine learning regression technique (as shown in Fig. 2).

The regression algorithm [31] is trained with specified training sets. Every training set corresponds to a specific user. The derived model is used to determine in real time the anxiety level of the user measured.

The bio-feedback application has a graphical interface that appears either on screen as desktop application or embedded in a web-based environment. The graphical part of the application receives the result of measurement classification and displays a visualised result on the computer screen where the user can recognize his/her anxiety states through a chromatic code (red for anxiety, green for relaxation, orange for moderate anxiety) besides to a percentage indication of

his/her anxiety levels. Every time the subject reaches high anxiety levels, the application encourages him/her to use the diaphragmatic breathing. This is a bio-feedback technique well known to psychologists and can be applied easily in almost every condition. This technique tries to reduce the intensiveness of stressful emotions. The application also displays a phrase chosen by the person who is participating in the measurement, in collaboration with a specialist psychologist who is supporting this research as a consultant. This phrase is associated with encouraging thoughts that may help the person who is measured to relax. Furthermore, there is a monitoring application that receives and displays the anxiety levels of people set in a group for overview by a teacher or by an intelligent tutor agent.

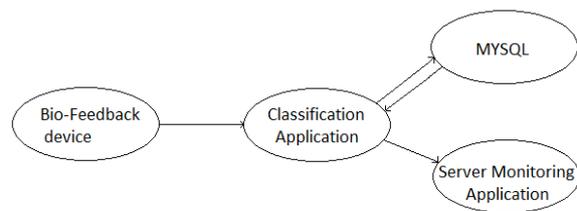


Fig 2. Bio-feedback System Architecture

V. EDUCATIONAL ACTIVITIES SUPPORTED BY E-BIO-FEEDBACK DEVICE

Exploring the various activities taken place, the support of the bio-feedback device was revealed. These activities were synchronous at the same place of a classroom or from a distance using open source videoconference tools like OpenMeetings (<http://openmeetings.apache.org/>) [32].

In these activities the bio-feedback device was measuring three kinds of bio-signals GSR, skin temperature and HR. Several tests were run measuring its usability and user interaction satisfaction. The instruments used for the device evaluation were the Questionnaire for User Interaction Satisfaction (QUIS) [33] as well as the system usability scale (SUS) [34]. In all the tests the majority of the results were positive. These activities are listed in Table 1 (Activities):

TABLE I. ACTIVITIES

Participants	Place	Level of anxiety awareness
10 Students (MSc in ICT in Education)	Synchronous at the same place	<ul style="list-style-type: none"> One to One (self-awareness) Many to One (teacher awareness about students' anxiety level) Many to Many (everyone is aware about others anxiety level)
	Synchronous from distance using videoconferencing	<ul style="list-style-type: none"> One to One (self-awareness), Many to One (teacher awareness about students' anxiety level)
10 Students (MSc in ICT in Education)	Synchronous at the same place	<ul style="list-style-type: none"> One to One (self-awareness), Many to One (teacher awareness about students' anxiety level)

<i>Participants</i>	<i>Place</i>	<i>Level of anxiety awareness</i>
		<ul style="list-style-type: none"> • Many to Many (everyone is aware about others anxiety level)
	Synchronous from distance using videoconferencing	<ul style="list-style-type: none"> • One to One (self-awareness), • Many to One (teacher awareness about students' anxiety level)
	Synchronous from distance in a 3D CVE	<ul style="list-style-type: none"> • One to One (self-awareness) • Many to Many (everyone is aware about others anxiety level)
4 Students (Dentists)	Synchronous at the same place	<ul style="list-style-type: none"> • One to One (self-awareness)
14 Students (MSc in Information Systems)	Synchronous at the same place	<ul style="list-style-type: none"> • One to One (self-awareness), • Many to One (teacher awareness about students' anxiety level)

In all described activities the participants were connected to the bio-feedback device. The anxiety measurements were taking place in real time. Every student was watching his/her online measurements. The teacher was watching his/her classroom anxiety levels on the monitoring application. In the cases of first and second activity the students were separated into groups of two or three members. The schedule and the deliverables of each activity were clearly stated at the beginning. Each group had identical assignments. The members of each group were chosen by the students themselves. Every group member could his/her own anxiety levels as well his/her group mates' measurements.

VI. PROPOSED USE CASES THAT HIGHLIGHT BIO-FEEDBACK DEVICE USAGE

This bio-feedback system can be widely used in online courses such as a Small Private Online Course (SPOC), to detect students' emotions at a distance and to keep the instructors informed about students' affective state. Many of the online courses offered at San Diego State University (California) have synchronous online meetings. Due to various reasons, many of the students choose not to use audio or video in online meetings. They interact with the instructor and peers through typing. Having a bio-feedback device will enrich the online classrooms and help to keep the students engaged in the learning process. Continuous and increasing exploration of the complex set of parameters surrounding online learning reveals the importance of the emotional states of learners and especially the relationship between emotions and effective learning (e.g. [35]). Research ([36]) also demonstrates that a slight positive mood does not just make you feel a little better but also induces a different kind of thinking, characterized by a tendency towards greater creativity and flexibility in problem solving, as well as more efficiency and thoroughness in decision making.

VII. CONCLUSIONS

For a person who participates in challenging and complex activities that require extra effort, it is very likely to experience strong emotions such as anxiety. The anxiety

regulation may be an important support to decrease high anxiety or increase anxiety levels in cases of relaxation. Therefore anxiety awareness can be a very useful tool to support student engagement, which in many cases is a key factor for better performance. So far the results of this device evaluation are encouraging. Especially in distance learning activities it would be useful to reduce the "didactical distance" between the trainer and the trainees.

VIII. FUTURE WORK

We aim to apply the bio-feedback system in different situations such as autonomous and collaborative learning activities. In addition the bio-feedback system could be used in work environment and in daily activities in a gamified approach so as to face daily anxiety. We aim to have many more participants in future testing and implementation. This will generate a better assessment of the developed system applying usability, ease of use, and usefulness evaluation tests.

Acknowledgement

This paper is supported by the Oriental Scholar program of Shanghai Municipal Education Commission (TPKY052WMJ).

REFERENCES

- [1] Shen, L., Wang, M. J., & Shen, R. (2009). Affective e-Learning: Using "Emotional" data to improve learning in pervasive learning environment. *Educational Technology & Society*, 12(2), 176-189. (11-20%)
- [2] D'Mello, S. K. & Graesser, A. C. (2012). Dynamics of Affective States during Complex Learning, *Learning and Instruction*, 22, 145-1.
- [3] Pekrun, R. & Stephens, E. J. (2010). Achievement Emotions. A control-value approach. *Social and Personality Psychology Compass*, 4, 238-255.
- [4] Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18, 315-341.
- [5] Baker, R., D'Mello, S., Rodrigo, M., & Graesser, A. (2010). Better to be frustrated than bored: the incidence and persistence of affect during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4), 223e241.
- [6] Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-adapted Interaction*, 19(3), 267e303.
- [7] Forbes-Riley, K., & Litman, D. (2010). Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech and Language*, 25(1), 105e126.
- [8] Rodrigo, M., & Baker, R. (2011). Comparing the incidence and persistence of learners' affect during interactions with different educational software packages. In R. Calvo, & S. D'Mello (Eds.), *New perspective on affect and learning technologies* (pp. 183e202). New York: Springer
- [9] Woolf, B., Bursleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: recognizing and responding to student affect. *International Journal of Learning Technology*, 4(3/4), 129e163
- [10] Lopatovska, I., & Arapakis, I. (2010). Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction. *Information Processing and Management*, doi:10.1016/j.ipm.2010.09.001
- [11] Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29 (3), 241-250.

- [12] Arapakis, I., Konstas, I., & Jose, J. M. (2009). Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In Proceedings of the seventeen ACM international conference on multimedia (pp. 461–470). USA: Springer.
- [13] Lopatovska, I. (2009). Emotional aspects of the online information retrieval process. Ph.D. Thesis. Rutgers: The State University of New Jersey.
- [14] Fasel, B., & Luetttin, J. (2003). Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(1), 259–275 (iDIAP-RR 99-19).
- [15] Jaimes, A., & Sebe, N. (2007). Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1–2), 116–134.
- [16] Pantic, M., & Rothkrantz, L. J. (2003). Toward an affect-sensitive multimodal human–computer interaction. *Proceedings of the IEEE*, 91(9), 1370–1390.
- [17] Breazeal, C. (2001). *Designing social robots*. Cambridge, MA: MIT Press.
- [18] Chan, C. H., & Jones, G. J. F. (2005). Affect-based indexing and retrieval of films. In Proceedings of the 13th annual ACM international conference on multimedia (pp. 427–430). New York, NY, USA: ACM.
- [19] Hanjalic, A., & Xu, L.-Q. (2005). Affective video content representation and modeling. *Multimedia, IEEE Transactions on*, 7(1), 143–154.
- [20] Boone, R. T., & Cunningham, J. G. (1998). Children’s decoding of emotion in expressive body movement: The development of cue attunement. *Developmental Psychology*, 34(5), 1007–1016.
- [21] de Meijer, M. (2005). The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*, 13(4), 247–268.
- [22] Wallbott, H. G. (1998). Bodily expression of emotion. *European Journal of Social Psychology*, 28(6).
- [23] Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., et al (2007). Multimodal emotion recognition from expressive faces, body gestures and speech. *Artificial intelligence and innovations 2007: From theory to applications* (Vol. 247, pp. 375–388). Berlin, Heidelberg: Springer-Verlag.
- [24] Castellano, G., Villalba, S. D., & Camurri, A. (2007). Recognising human emotions from body movement and gesture dynamics. In Proceedings of the 2nd international conference on affective computing and intelligent interaction (pp. 71–82). Berlin, Heidelberg: Springer-Verlag.
- [25] Chen, F.-S., Fu, C.-M., & Huang, C.-L. (2003). Hand gesture recognition using a real-time tracking method and hidden markov models. *Image and Vision Computing*, 21(8), 745–758.
- [26] Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2), 256–274.
- [27] Wilhelm, F. H., Pfaltz, M. C., & Grossman, P. (2006). Continuous electronic data capture of physiology, behavior and experience in real life: Towards ecological momentary assessment of emotion. *Interacting with Computers*, 18(2), 171–186.
- [28] Lykken, D. T., Venables, P. (1971): Direct measurement of skin conductance: A proposal for standardization. *Psychophysiology*, 8, 656–672. Designated a Citation Classic, Institute for Scientific Information.
- [29] Challoner A. V. J. (1979). Photoelectric plethysmography for estimating cutaneous blood flow Non-Invasive Physiological Measurements vol 1 ed P Rolfe (London: Academic) pp 125–51.
- [30] Trafford, J., Lafferty, K. (1984). What does photoplethysmography measure? *Med Biol Engl Comput* ; 22: 479 –80.
- [31] Fisher, R.A. (1922). "The goodness of fit of regression formulae, and the distribution of regression coefficients". *Journal of the Royal Statistical Society* (Blackwell Publishing) 85 (4): 597–612. doi:10.2307/2341124
- [32] H. Apostolidis, P. Stylianidis, & Th. Tsiatsos. Augmenting the Educational Process Using a Prototype Bio-Feedback Device for Anxiety Awareness. In Karagiannidis Ch. Research on e-Learning and ICT in Education. 2014.
- [33] Chin, J.P., Diehl, V.A., & Norman, K.L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. CHI '88 Proceedings of the SIGCHI conference on Human factors in computing systems. 213-218.
- [34] Brooke, J. (1996). SUS – A quick and dirty usability scale. Retrieved September 6, 2013 from <http://hell.meiert.org/core/pdf/sus.pdf/>.
- [35] Kort, B., Reilly, R., & Picard, R. W. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. Proceedings of the IEEE International Conference on Advanced Learning Technologies, Los Alamitos, CA: IEEE Computer Society Press, 43-46.
- [36] Isen, A. M. (2000). Positive affect and decision making. In M. Lewis & J. Haviland (Eds.), *Handbook of emotions* (pp. 720), Guilford, New York: The Guilford Press.
- [37] Laure, D. (2012). “Heart rate measuring using mobile phone’s camera.” in Proceedings of the 12th Conference of Open Innovations Association FRUCT and Seminar on e-Travel. Oulu, Finland. St.-Petersburg: SUAI, pp. 272–273.
- [38] Poh, Ming-Zher, McDuff, D. J., & Picard, R.W. (2010). Non-contact automated cardiac measurements using video imaging and blind source separation in *Optics Express* 18 (2010): 10762.

Hippokratis Apostolidis is currently a Ph.D. candidate of the laboratory of “Information and Communication Technologies in Education” of Computer Science Department at Aristotle University of Thessaloniki. He obtained his Bachelor from the school of Mathematics of Aristotle University of Thessaloniki, his Master’s degree in Computer Science from the Computer Science department of University of Essex and his Master’s degree in ICT from Computer Science department of Aristotle University of Thessaloniki. He is working as programmer – analyst in banking applications. He was an Associate Professor in Technological Institution of Thessaloniki (ATEITH). He is a member of the research team of the European funded project of ASPAD. His research interests are on the scientific fields of affective computing, machine learning, cognitive training, robotics in education, augmented reality in education and computer supported collaborative learning (CSCL).

Thrasylvoulos Tsiatsos is currently an Assistant Professor in the Department of Informatics of Aristotle University of Thessaloniki and a member of the Multimedia Lab. He obtained his Diploma, his Master’s Degree and his PhD from the Computer Engineering and Informatics Department of Patras University (Greece). His research interests include: Networked Virtual Learning Environments; Internet Technologies in Education; Assessment in Online Educational Environments; Game Based Learning Using New Technologies; Mobile Learning; Cognitive Training Using Information and Communication Technologies; Development of Bio-Feedback Interfaces to Support in The Educational Process and Learning; Design and Development Internet Environments for the Management and Support of Education and Training; Open and Distance Education Using Multimedia and Internet Technologies; Collaborative learning and Distance learning. He has published more than 160 papers in Journals and in well-known refereed conferences and he is co-author in 3 books. He has been a PC member and referee in various international journals and conferences. He has participated in R&D projects such as OSYDD, RTS-GUNET, ODL-UP, VES, ODL-OTE, INVITE, EdComNet, VirRAD, SAPSAT, E-internationalization for Collaborative Learning (EICL) and Education of oreign and Repatriated Students (NSRF – National Strategic Reference Framework, 2007–2013). Also he is member of IEEE and Technical Chamber of Greece.

Minjuan Wang is an oriental scholar at Shanghai International Studies University (China), a professor of Learning Design and Technology at San Diego State University (USA), and a Program Manager for the Chancellor’s office of California State University (USA). Her research specialties focus on the sociocultural facets of online learning, and the design and development of mobile and intelligent learning. She has published peer-reviewed articles in *Educational Technology Research and Development*, *Computers and Education*, *Educational Media International*, *TechTrends*, and the *British Journal of Educational Technology*. She has also published book chapters on engaged learning in online problem solving, Cybergogy for interactive learning online, informal learning via the Internet, and effective learning in multicultural and multilingual classrooms. Address for correspondence: Dr MinjuanWang, 5500 Campanile Dr. PSFA 315, SDSU, San Diego, CA 92182-4561. Tel: 619-5943878 Email:mwang@mail.sdsu.edu.

Bio-inspired algorithms for attack of block Ciphers

T. Mekhaznia
 LASIC Laboratory
 University of Batna, Algéria
 t.mekhaznia@univ-tebessa.dz

A. Zidani
 Department of Computer Sciences
 Université of Batna, Algeria
 zidani@free.fr

Abstract— Block ciphers cryptosystems are an assurance to the security and safety of data. They constitute a hard task for cryptanalysis attacks due to the nonlinearity of their structure. This problem is viewed as NP-Hard; various research efforts has been accomplished in its resolution but numerous attacks results still insufficient especially when handling wide instances due to resources requirement which increase with the size of the problem. On another way, bio-inspired intelligence algorithms are a set of approaches that mimics the real nature swarms of insects for solving optimization problems; they are widely used in computational research application and characterized by their fast convergence with reasonable time consumption. The scope of this paper is to provide more detailed study about the performance of two bio inspired algorithms BAT algorithm and Wolf Pack Algorithm for cryptanalysis of some basic block ciphers within limited computer resources. Experiments were accomplished in order to study the effectiveness of the used algorithms in solving the considered problem.

Keywords— Cryptanalysis, block ciphers, bio inspired algorithms.

I. INTRODUCTION

Cryptography refers to the science of protection of information and preventing access even from malicious ones and developing avenues of encryption methodologies. It involves an input data called a *plaintext* and a small amount of information: a *key*, in order to customize an obfuscated an output data called a *ciphertext*, a unintelligible information to all not intended parties. The set of plaintext, ciphertext, keys and encryption algorithms is called a cryptosystem. Cryptanalysis denotes to is the art of studying encryption information, ciphers and related concepts in order to detect their weakness and attempt to extract parts of corresponded plaintext without knowing the secret data which normally used for decryption such keys or algorithms. The use of cryptanalysis on ciphertexts is called an *attack*; if a plaintext can be restored, the attack is successful. So, the design of complex cryptosystems seems a challenge against research in communication security, particularly with the increasing of amount of data transfer through public networks. The goal of cryptanalysis is based upon the knowledge of cryptosystems features such fragments of

cipher and plain data, encryption algorithms and if possible, language characteristics. This information favorite right attacks [1]; its aims is to measure cryptosystems strength and therefore, help the researchers to perform more robust algorithms for upcoming times.

Cryptanalysis uses several techniques of attacks such *linear cryptanalysis* [2], *differential cryptanalysis* [3], *integral cryptanalysis* or *slide attack* [4] which are based, in general on known or chosen plaintexts. These techniques are able to break various ciphers, nevertheless, and given their reduced setting, they remain ineffective against modern cryptosystems. The *brute force* is a common attack; it tries the 2^b possibilities of b -length key within the search space to find the right key. It's a successfully way to break ciphertexts but need enough resources and so, has less success in practice.

Block ciphers are algorithms for symmetric key encryption scheme which operate on large blocks of data. They built based on nonlinearity and low autocorrelation and characterized by their simplicity of implementation, high speed of encryption [5] and resistance against various attacks [6].

Research in cryptanalysis is, actually intended to explore heuristic techniques based on bio-inspired intelligence which are general purpose approaches that proved their efficiency to enable complex search spaces. Bio-inspired intelligence algorithms are a well-known paradigm that successfully used for implementation of powerful tools for solving tough problems [7] and especially cryptanalysis problems with reasonable amount of resources consumption [8]. Various works [9] [10] shown that algorithms based swarm intelligence have a successful potential to handle wide instances and may be adapted to produce approximate solutions for a large variety of optimization problems. They use intelligent system that offers an independence of movement of agents and tend to replace the preprogramming and centralized control. In last years, many of such algorithms were emerged [11].

The aim of the presented paper is to investigate the way in which the bio-inspired algorithms and, in particular BAT and WPA algorithms can be used for attack on encryption keys of Block ciphers. This technique provides a successful

way of automated cryptanalysis by using limited parameters without any need of initial approximation to unknown parameters. Also, it has been successfully applied in a wide range of research application areas. It is proved that it gets better results in a faster and cheaper way.

II. LITERATURE REVIEW

Research in swarm intelligence and, bio-inspired intelligence in particular, has used in a significant part of cryptanalysis attacks in recent decays. Various works [12] [9] [13] shown that algorithms based bio-inspired intelligence have a successful potential in handling wide instances of variety of optimization problems. In their works, Laskari et al. [14] [15] used a variety of computational algorithms in resolution of cryptographic problems and demonstrate their effectiveness in cryptosystems security. On another way, bio-inspired algorithms has used for attack of block ciphers such 4DES and DES [16] [17] [18] [19]; experiments revealed most bits of used keys and showed that these algorithms may be a powerful tool in cryptanalysis of such problem.

III. BLOCK CIPHERS

Block ciphers [20] are iterated encryption structure which produces fixed length blocks of *ciphertext* from similar blocks of *plaintext* by a sequential r times repetition of a nonlinear complex transformation (called *round function*) based on several substitution and permutation on block-bits.

In order to produce a fixed size ciphertext C from an identical size plaintext M , The block cipher algorithm proceeds for each bloc of n -bits of M , r iterations of the round function f_i using previous encrypted block and a subkey i . Initially, each block is split into two halves L_0 and R_0 of $n/2$ bits each. At iteration i , the round function is applied to one half using a subkey, the output is exclusive-oriented with the other half. The two halves are then swapped as shows in following relation:

$$(L_i, R_i) = (R_{i-1}, L_{i-1} \oplus f_i(R_{i-1}, k_i)) \quad (1)$$

where f_i ($i > 0$), a nonlinear function usually represented as a substitution boxe (called *sboxe*). It substitutes an input of n bits size with an output of m bits size ($m < n$). This ensures that all subsequent blocks are different.

The advantage of the algorithm is that the encryption and decryption functions are identical. To allow a unique decryption, the encryption transformation must be a bijection, defining one-to-one on n -bits of each encrypted block. So, to reverse a round, it is only necessary to apply the same transformation again, which will cancel the changes of the binary operation XOR.

Block ciphers become a basis component in many encryption schemes Blowfish [21]. LUCIFER [22], CAST [23], IDEA [24], AES [25], RC5 [26].

IV. BLOCK CIPHERS CRYPTANALYSIS

To evaluate the security of block ciphers, we assume that the attacker has access to all ciphered data and some details about the encryption algorithm. So, the cryptanalysis is limited to the *ciphertext attack only* which seems the most challenging kind of attacks [27]. Under this assumption, the attacker generates several chosen keys and proceeds to decryption. Upon these consideration, we show that blocks with reduced data are insecure (e.g, a key of 10 bits, such SDES algorithm, implying 2^{10} different ciphertexts) which can be easily attacked by *exhaustive search* or *frequency analysis*. The blocks security can be increased by using wide block length. Although, it becomes unfeasible for an attacker to try every possible key until the desired plaintext is found for blocks more than 56 bits such DES algorithm, that needs 2^{56} combinations when using exhaustive search, 2^{43} *known plaintexts* with *linear cryptanalysis* (Matsui, 1994) and 2^{47} chosen plaintexts when using *differential cryptanalysis* [28]. However and in order to avoid exhaustive search, actual research in cryptanalysis tends toward and especially, bio-inspired algorithms which have been found efficient in resolution of this kind of problems.

The cryptanalysis by using bio-inspired algorithm consists on searching keys with the maximum of correct bits in regard of the encryption key. Since, the encryption key isn't available as an input data, the result is evaluated according to the produced readable plaintext. A readable plaintext should include some properties of the language with which it was written.

A. Language properties

The most important property of a given language is the distribution statistics of its characters on written texts. The frequency analysis is used to define the quality of plaintexts produced by an attack process. *Letter frequency distribution* isn't uniform and different from one language to another. The general order of character occurrence is based upon various studies on language spelling and syntax by using various sets of text (articles, books, newspaper, media, etc.) called *Corpus* [29] [30] and illustrated on tables called '*letter frequency table*' [31] [32].

The *letter frequency analysis* is the study of the occurrence of single or group of letters and determining at which frequency a letter of the plaintext occurs within the corresponding ciphered text. This fact provides indications about ciphertext language and helps in recover certain of its letters, especially the most frequent.

B. Fitness function

The fitness function measures the difference between the letter frequency of produced plaintext with letter frequency of standard language (such illustrated in table 1). A close difference denotes an acceptable solution. In literature, the fitness function has been proposed under various combination schemes [33] [34] [35] [36] [37] [38] [39] [17].

The most commonly used is given by following equation:

$$F(k) = \alpha \sum_{i=1}^{26} |D(i) - C(i)|^u + \beta \sum_{i,j=1}^{26} |D(i,j) - C(i,j)|^b. \quad (2)$$

where D , C denotes respectively known language statistics and decrypted text statistics. u , b : denotes respectively 1-gram and 2-gram statistics, α and β (with $\alpha+\beta=1$) are weights assigning different priorities to 1-gram and 2-gram and k , the key used in decryption process.

Table 1 Frequency table for most English and French Corpus

Letter frequency (%)				
	Thomas Tempé [1]	Concise OD [2]	Corpus Français [3]	Leipzig Corpora [4]
A	7.246	8.167	9.38	8.55
B	0.855	1.492	1.54	1.6
C	3.094	2.782	1.49	3.16
D	3.481	4.253	4.70	3.87
E	13.98	12.702	10.15	12.1
F	1.012	2.228	2.03	2.18
G	0.822	2.015	2.86	2.09
H	0.699	6.094	2.09	4.96
I	7.144	6.966	5.82	7.33
J	0.517	0.153	0.61	0.22
K	0.046	0.772	3.14	0.81
L	5.177	4.025	5.28	4.21
M	2.816	2.406	3.47	2.53
N	6.732	6.749	8.54	7.17
O	5.121	7.507	4.48	7.47
P	2.867	1.929	1.84	2.07
Q	1.292	0.095	0.02	0.1
R	6.218	5.987	8.43	6.33
S	7.542	6.327	6.59	6.73
T	6.874	9.056	7.69	8.94
U	5.988	2.758	1.92	2.68
V	1.545	0.978	2.42	1.06
W	0.108	2.360	0.14	1.83
X	0.367	0.150	0.16	0.19
Y	0.292	1.974	0.71	1.72
Z	0.129	0.074	0.07	0.11

V. SWARM INTELLIGENCE HEURISTICS

Based on natural systems that dedicated to solve complex problems [40], swarm intelligence heuristics (*SI*, in short) is an idea introduced by [41] and denoted sets of entities interacting and exhibiting a collective intelligence behavior

[1] <http://web.archive.org/web/20080213211515/http://gpl.insa-lyon.fr/Dvorak-Fr/CorpusDeThomasTemp%C3%A9>.

[2] <http://en.algorithmy.net/article/40379/Letter-frequency-English>.

[3] http://wortschatz.uni-leipzig.de/ws_fra.

[4] <http://corpora.informatik.uni-leipzig.de/>

according to natural rules in order to accomplish a common task. *SI* heuristics uses a basic population of individuals which represents candidate solutions. They are doted of a capability to act in a synchronized and decentralized manner without need of coordination [42]. This principle allows handling very large space of solutions but no guarantee of an optimal solution is ever found given that, the research becomes useless, if in a exploration space, a cross between the local and global solution occurs [43].

Bio-inspired heuristics, a branch of swarm intelligence heuristics, are adaptive strategies that emerges though the interaction and cooperation of real world insect swarms and used as a problem solving tool. This principle is used to produce algorithms able to resolve complex tasks without centralized control.

VI. BAT ALGORITHM

BAT Algorithm [44], is a metaheuristic population based approach, inspired from the hunting behavior of bats which uses the echolocation to sense and discriminate surrounding objects.

In their flying and, in order to avoid different obstacles and locate prey, Bats emit loud sonar [45] throughout their environment; the echo that bounces back allows identifying kinds of surrounding objects. Studies [22] show that the loudness of emitted pulse varies from lowness rate with a long duration of sounds when exploring hunt area to loudest with a decreasing duration of sounds when homing toward prey.

For simulation, we use virtual bats with numerous positions and velocities in a virtual space search that evaluate according to following steps:

- In a search space, and at time t , each bat i has a position x_i^t and a velocity $v_i^t \in \mathbb{R}^n$.
- In their randomly fly with a constant velocity v , each bat i emit a uniformly pulse frequency f_{min} .
- At the perception of a prey, above parameters are adjusted depending on the distance to prey according to following relations:

$$f_i = f_{min} + \beta(f_{max} - f_{min}) \quad (3)$$

$$v_i^{t+1} = v_i^t + f_i(x_i^t - x_g^t) \quad (4)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (5)$$

where β a random vector distributed in range $[0,1]$, x_g the position of best bat of swarm.

VII. WOLF PACK ALGORITHM

Wolf Pack Algorithm (*WPA*, in short) is a new swarm intelligence algorithm inspired by hunting behavior of artificial predators [46] based on *Wolf Pack Search* [47]. It consists basically in making predators (wolves) hunt, find the trail of prey and capture it under the command of a lead wolf.

WPA is an efficient and robust search method which permits to avoid local optima. It used to approximate solutions to real-word optimization problems.

The Wolf Pack is built as a social work division contains a *leader-wolf* which is the strongest one; it's responsible of the pack management. Its decisions always based upon the activity of other animals in surrounding environment: prey, wolves of pack and other predators. The pack includes two classes of wolves: *elite-wolves* which move independently around the environment and adjust their direction according to the concentration of prey's smell. When a prey is discovered by an elite-wolf, it will howl. The perception of voice by the lead-wolf announces the attack by *furious-wolves* which get close to howl position. After capturing a prey, it assigned in an order from the strong to weak. This fact causes the dead of weak wolves for lack of food which enhanced the pack and keeps it strongly at any time.

The Wolf Pack algorithm is accomplished by following steps:

- In a search space \mathbb{R}^n , each artificial wolf i represent a basic solution of the problem, has a position x_i . Initially, wolves are randomly distributed in space.
- At each instant t , the wolf i perform a move from position x_i^t to position x_i^{t+1} . The choice of next position is updated according to the following equation:

$$x_i^{t+1} = x_i^t + \lambda |x_g^t - x_i^t| \quad (5)$$

where λ a random vector distributed in range $[-1,1]$, x_g the position of the lead-wolf.

- After a fixed number of iterations, which corresponds to a scouting phase, the wolf of the best solution became a lead wolf; a certain number of weak wolves (bad solutions) will be deleted and replaced by a new generation of random wolves.

VIII. PROBLEM FORMULATION

In cryptanalysis of block ciphers, each bat/wolf i represent a basic solution of the problem that corresponds to a decryption key k , a vector of n bits; each bit x represents a bat/wolf position in their exploration. At each move from position x_i^t to position x_i^{t+1} , the bat/wolf i perform a decryption key $k_{x^t, x^{t+1}}$ obtained by swapping the bits x^t and x^{t+1} according respectively to equations (5) and (6) in case of BAT or WPA algorithms. The performance of each position corresponds to the quality of plaintext obtained using the decryption key $k_{x^t, x^{t+1}}$ according to equation (2). The process will be stopped after a fixed number of iterations or if no improvement in solution.

Algorithm 1 outlines the main steps of BAT algorithm applied to the cryptanalysis of block ciphers.

Algorithm 1. BAT

Input: Cipher_n, BatNumber, S_{BatNumber}, $k_{\text{BatNumber}}, k^*, \beta$
Output: S*
 Generate k_i, f_i, v_i, c_i (i=1.. BatNumber)
 Evaluate S_i , (i=1.. BatNumber)
 $S^* \leftarrow \min(S_i, i=1.. \text{BatNumber})$
While not (*exit criterion*)
 For $j \leftarrow 1$ to *KeySize* **do**
 Pick random numbers: $\beta \in [0,1]$
 For each bat i **do**
 $f_i = f_{\min} + \beta(f_{\max} - f_{\min})$ with $f_i \in \{f_{\min}, f_{\max}\}$
 $v_i \leftarrow v_i + |f_i(k_i - k^*)|$ with $v_i \in \{v_{\min}, v_{\max}\}$
 if ($v_i > 0.5$) **then** $k_i \cdot \text{bit}_j \leftarrow 1$
 else $k_i \cdot \text{bit}_j \leftarrow 0$
 endif
 If ($S_{ki} > S^*$) **then** $S^* \leftarrow S_i$ **Endif**
 Endfor
EndWhile
 Report S*

Algorithm 2 illustrated WPA in cryptanalysis of bloc ciphers:

Algorithm 2. WPA

Input: Cipher_n, WolfNumber, S_{WolfNumber},
 $k_{\text{WolfNumber}}, k^*, it, \lambda$
Output: S*
 Generate k_i , (i=1.. WolfNumber)
 Evaluate S_i , (i=1.. WolfNumber)
 $S^* \leftarrow \min(S_i, i=1.. \text{WolfNumber})$
While not (*exit criterion*)
 $Iter_{\text{scoot}} \leftarrow 0$
 While ($Iter_{\text{scoot}} < It$)
 For $j \leftarrow 1$ to *KeySize* **do**
 Pick random numbers: $\lambda \in \{0,1\}$
 For each wolf i **do**
 $k_i \cdot \text{bit}_j \leftarrow k_i \cdot \text{bit}_j + \lambda |k^* \cdot \text{bit}_j - k_i \cdot \text{bit}_j|$
 If ($S_i > S^*$) **then** $S^* \leftarrow S_i$ **Endif**
 Endfor
 Update $Iter_{\text{scoot}}$
 Endfor
 Delete worst-wolf w / ($S_w = \max(S_i, i=1.. \text{WolfNumber})$)
 Generate a new random w
EndWhile
 Report S*

IX. EXPERIMENTATION AND RESULTS

In order to outline the performance of the proposed algorithms, some experiments have been conducted on a set of sample binary texts of 800 to 12000 bits (100 to 1500 ASCII characters) extracts from ICE [48]. Encryption methods used are: 4DES [49], FEAL-4 [50] and RC5 [26]

cryptosystems using four rounds encryption scheme. Each key is a binary vector of 64 bits (in case of 4SDES and FEAL-4) and 40 bits in case of RC5. The used algorithm is coded on Matlab 2.14 and performed on a CPU 3.2 Ghz.

The results obtained after carrying the mentioned experiments are illustrated in figures 1 and figure 2 below which show the average percentage of recovered bits of encryption key for each algorithm when using a fixed processing time of 150 seconds.

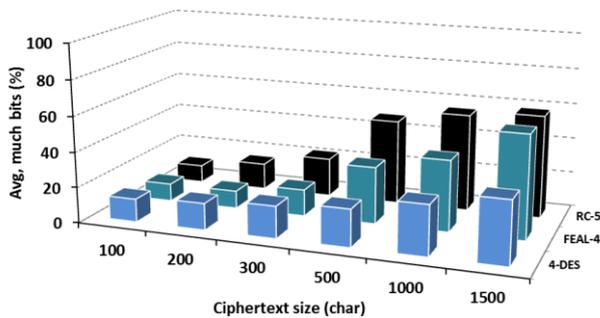


Fig. 1. Performance evaluation of BAT Algorithm

In this experiment, we show that both algorithms perform an average of more than 60% of bit-keys with reasonable amount of time processing. Although, the WPA algorithm outperforms significantly than BAT algorithm. Also, the performance of both BAT and WPA is particularly noticeable when using texts of more than 500 characters.

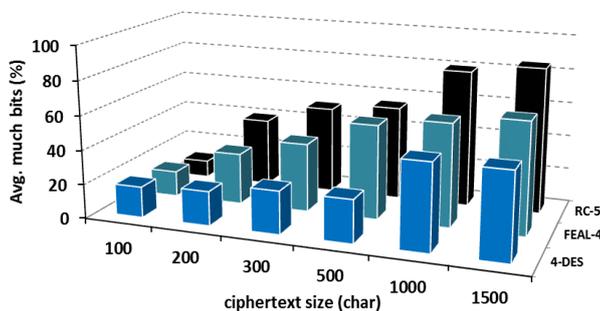


Fig. 2. Performance evaluation of WPA Algorithm

X. CONCLUSION

Bio inspired algorithms are denote as a new revolution of efficient research strategy in resolution of complex combinatory tasks such cryptanalysis problem. In this paper, a comparative study of BAT algorithm and WPA algorithm for cryptanalysis of some variants of block ciphers was conducted. The experiments show that the algorithms can be successfully applied in resolve of such problem. The produced results show that both algorithms allow locating more than 60% of bits-key with acceptable resource consumption. The tests were operates on a reduced space of data, however, the results presented can be improved by the well choice of problem factors such environment parameters, ciphered data and languages statistics. In addition, the study may open avenues to investigate the effectiveness of other evolutionary computation techniques for further attacks of more complicated cryptosystems.

REFERENCES

- [1] K. V. S. Rao, M. R. Krishna and B. Babu, "Cryptanalysis of a Feistel Type Block Cipher by Feed Forward Neural Network Using Right Sigmoidal Signals," *International Journal of Soft Computing*, pp. 131-135, 2009.
- [2] M. matsui, "Linear Cryptanalysis Method for DES Cipher," *Springer*, pp. 386-97, 1993.
- [3] E. Biham and A. Shamir, "Differential cryptanalysis of the data encryption standard," *Springer-Verilag*, pp. 8-9, 1993b.
- [4] A. Biryukov and D. Wagner, "Slide Attacks," in *Fast Software Encryption*, Rome, 1999.
- [5] S. A. Alomari, P. Sumari and A, "Spiking Neurons with ASNN Based-Methods for the Neural Block Cipher," *International journal of computer science & information Technology*, vol. 2, no. 4, pp. 138-148, 2010.
- [6] R. Singh and D. B. Ojha, "An Prdeal random data Encryption Scheme," *International Journal of Engineering Science and technology*, vol. 2, no. 11, pp. 6349-6360, 2010.
- [7] A. Gandomi and A. Alavi, "Stage Genetic Programming: A New Strategy to Nonlinear System Modeling," *Information Sciences*, vol. 181, no. 23, pp. 5227-5239, 2011.
- [8] Gherboudj A., Chikhi S., "BPSO Algorithms for Knapsack Problem," *Communications in Computer and Information Science*, pp. V162:217-227, 2011.
- [9] C. Blum and X. Li, "Swarm Intelligence in Optimization," *Natural Computing series*, pp. 43-85, 2008.
- [10] F. T. S. Chan and M. K. Tiwari, *Swarm Intelligence, Focus on Ant and Particle Swarm Optimization*, Vienne, Austria: I-Tech Education and Publishing,

- 2007.
- [11] G. S. Sharvani, N. K. Cauvery and T. M. Rangaswamy, "Different Types of Swarm Intelligence Algorithm," in *Int. Conf. on Advances in Recent Technologies in Communication and Computing*, Kottayam, Kerala, India, 2009.
- [12] T. S. C. Felix and K. T. Manoj, *Swarm Intelligence: Focus on Ant and Particle Swarm Optimization*, Vienna: Itech Education and Publishing, 2007.
- [13] A. Dadhich, A. Gupta and S. Yadav, "Swarm Intelligence based linear cryptanalysis of four-round data Encryption Standard algorithm," in *Int. Conf. on Issues and Challenges in Intelligent Computing Techniques*, Ghaziabad, 2014.
- [14] E. C. Laskari, G. C. Meletiou, Y. C. Stamatiou and M. N. Vrahatis, "Cryptography and cryptanalysis through computational intelligence," *Studies in Computational Intelligence*, pp. 1-49, 2007b.
- [15] E. C. Laskari, G. C. Meletiou, Y. C. Stamatiou and M. N. Vrahatis, "Applying evolutionary computation methods for cryptanalysis of Feistel ciphers," *Applied Mathematics and Computation*, vol. 184, pp. 63-72, 2007a.
- [16] W. Shahzad, A. B. Siddiqui and F. Khan, "Cryptanalysis of Four-Rounded DES using Binary Particle Swarm Optimization," in *Genetic and Evolutionary Computation Conference*, NY, 2009.
- [17] Vimalathithan R., Valarmathi M. L., "Cryptanalysis of DES using Computational Intelligence," *European Journal Of Scientific Research*, pp. V55(2):237-244, 2011b.
- [18] G. A. Wafaa, N. I. Ghali, A. E. Hassanien and A. Abraham, "Known Plaintext Attack of DES-16 using Particle Swarm Optimization," in *Third World Congress of nature and Biologically Inspired Computing*, 2011.
- [19] S. Pandey and M. S. Mishra, "Particle Swarm Optimization in Cryptanalysis of DES," *Int. J. of Advanced Research in Computer Engineering & Technology*, vol. 1, no. 4, pp. 379-381, 2012.
- [20] H. Feistel, "Cryptography and computer privacy," *Scientific American*, vol. 228, no. 5, pp. 15-23, 1973.
- [21] B. Schneier, "Description of a New Variable-Length Key, 64-Bit Block Cipher Blowfish," *Fast Software Encryption, Cambridge Security Workshop Proceeding*, pp. 191-204, 1993.
- [22] J. R. Speakman and P. A. Racey, "The cost of being a bat," *Nature*, pp. 421-423, 1991.
- [23] M. A. Carlisle, "Constructing of Symmetric ciphers using the CAST design Procedure," *Designs, Codes, and Cryptography*, pp. 283-316, 1997.
- [24] X. Lai and J. L. Massey, "A proposal for a new block encryption standard," in *Workshop on the theory and application of cryptographic techniques on Advances in cryptology (EUROCRYPT)*, NY, 1991.
- [25] FIPS, "Announcing the Advanced Encryption Standard (AES)," Publication 197, U.S. DoC/NIST, 2001.
- [26] R. L. Rivest, "The RC5 Encryption Algorithm," *Fast Software Encryption*, pp. 86-96, 1995.
- [27] A. Biryukov and A. Kush, "From differential cryptanalysis to ciphertext-only attacks," *Lecture notes in Computer Sciences*, pp. 72-88, 1998.
- [28] E. Biham and A. Shamir, "Differential cryptanalysis of the full 16-round DES," *Advances in Cryptology*, pp. 487-496, 1993a.
- [29] L. Robert, *Cryptological mathematics*, The mathematical Association of America, 2000.
- [30] G. Nelson, S. Wallis and B. Aarts, *Exploring Natural Language*, John benjamins Publishing Company, 2002.
- [31] S. Singh, *The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography*, NY: Doubleday, 1999.
- [32] H. Beker and F. Piper, *Ciphers Systems: The Protection of Communications*, Northwood Books, 1982.
- [33] N. Nalini and G. R. Rao, "Attacks of simple block ciphers via efficient heuristics," *Information Sciences*, vol. 177, no. 12, pp. 2553-2569, 2007.
- [34] R. Spillman, M. Janssen, B. Nelson and M. Kepner, "Use of a genetic algorithms in the cryptanalysis of simple substitution ciphers," *Cryptologia*, pp. 31-44, 1993.
- [35] J. A. Clark, "Modern Optimization Algorithms for Cryptanalysis," in *Second Australian and New Zealand Conference on Intelligent Information Systems*, Brisbane, Qld, 1994.
- [36] J. P. Giddy and N. R. Savafi, "Automated cryptanalysis of transposition ciphers," *The Computer Journal*, pp. 429-436, 1994.
- [37] J. A. Clark and E. Dawson, "A parallel genetic algorithm for cryptanalysis of the polyalphabetic substitution cipher," *Cryptologia*, pp. 129-138, 1997.
- [38] M. D. Russel, J. A. Clark and S. Stepny, "Making the most of two heuristics: Breaking transposition ciphers with ants," in *Congress of Evolutionary Computation*, 2003.
- [39] M. F. Uddin and A. M. Youssef, "Cryptanalysis of simple substitution ciphers using particle swarm optimization," in *IEEE Congress on Evolutionary Computation*, Vancouver, CA, 2006a.
- [40] E. Bonabeau, M. Dorigo and G. Theraulaz, *Swarm Intelligence. From Natural to Artificial Systems*, Oxford: Oxford University Press, 1999.
- [41] G. Beni and J. Wang, "Swarm Intelligence in Cellular Robotic Systems," in *Advanced Workshop on Robots and Biological Systems*, Tuscany, Italy, 1989.
- [42] P. Tarasewich and P. R. McMullen, "Swarm Intelligence: Powers in numbers," vol. 45, no. 8, pp. 62-67, 2002.

- [43] J. Olamaei, T. Niknam and G. Gharehpetian, "Application of particle swarm optimization for distribution feeder reconfiguration considering distributed generators," *Applied Mathematics and computation*, vol. 201, no. 1, pp. 575-586, 2008.
- [44] X. S. Yang, "A New Metaheuristic Bat-Inspired Algorithm," in *Nature Inspired Cooperative Strategies for Optimization*, 2010.
- [45] D. Griffin, *Listening in the dark: the acoustic orientation of bats and men*, Cambridge, MA: yale University Press, 1958.
- [46] H. S. Wu and F. M. Zhang, "Wolf Pack Algorithm for Unconstrained Global Optimization," *Mathematical Problems in Engineering*, pp. 1-17, 2014.
- [47] C. Yang and J. Chen, "Algorithm of marriage in Hooney Bees Optimization Bosed on the Pack Search," in *International Conference on Intelligent Pervasive Computing*, Jeju, 2007.
- [48] N. Gerald, W. Sean and A. Bas, *Exploring Natural Language*, John Benjamins Publishing Company, 2002.
- [49] E. C. Laskari, G. C. Meletiou, Y. C. Stamatio and M. N. Vrahatis, "Evolutionary computation based cryptanalysis: A first study," *Nonlinear Analysis: Theory, Methods and Applications*, pp. e823-e830, 2005.
- [50] A. Shimizu and S. Miyaguchi, "Fast Data Encipherment Algorithm," *Lecture Notes in Computer Sciences*, 1987.

Influence of mesh quality and density on numerical calculation of heat exchanger with undulation in herringbone pattern

Václav Dvořák, Jan Novosád

Abstract— Research of devices for heat recovery is currently focused on increasing the temperature and heat efficiency of plate heat exchangers. The goal of optimization is not only to increase the heat transfer or even moisture but also reduce the pressure loss and possibly material costs. During the optimization of plate heat exchangers using CFD, we are struggling with the problem of how to create a quality computational mesh inside complex and irregular channels. These channels are formed by combining individual plates or blades that are shaped by molding, vacuum forming, or similar technology. Creating computational mesh from the bottom up manually is time consuming and does not help later optimization. The paper presents a comparison of results obtained from numerical simulations using meshes created by two different ways. The first way is creating the mesh manually. This method is quite slow and difficult. It is necessary to create new mesh for each variant of heat exchange surface. The second way is creating meshes based on dynamic mesh method provided by software Fluent. Creating of mesh by pulling is similar to the own production process, i.e. it is perpendicular to the plates. The paper discusses the differences in results of numerical simulations using meshes creating by different methods with various element size. It was found that generally finer meshes bring lower obtained efficiency and lower pressure loss.

Keywords— Dynamic mesh, heat exchanger, CFD.

I. INTRODUCTION

The development of recuperative heat exchangers in recent years focused on increasing efficiency. Another challenge is the development of so-called enthalpy exchangers for simultaneous heat and moisture transport, i.e. transport of both sensible and latent heat, as presented by Vít et al. in work [1].

To simulate a heat exchanger, we have to create a model and a computational mesh and use computational fluid dynamic (CFD) software. By assembling the heat exchanger, complicated and irregular narrow channels are created. These channels are split into small volumes (elements). Final mesh should be structured or unstructured with different element

size.

A lot of others researchers dealt with design and investigation of performance and pressure drop of plate heat exchangers based on numerical simulations. Most of them used the unstructured mesh for calculations.

Gherasim et al. in work [2] presented the comparison of various grids for plate heat exchanger modelled by tetrahedral mesh. In order to assess the influence of the grid resolution on the solution, five grids were created and tested by meshing the volumes with different interval sizes. The laminar and turbulent regimes were simulated. The evolutions of the average pressure and average temperature of the hot fluid over transversal sections along the length of the plate was investigated. In general, the differences between the series for the turbulent case are larger than those for the laminar case. It was founded that the two grids with smallest elements give very close results. In terms of temperature the obtained results were closed for grids with smaller elements. For the pressure, there were founded a quite large difference between the grid with smallest elements and the grid with the largest ones.

There are some next researchers who dealt with numerical simulations of plate heat exchangers with the chevron (undulated) profile. E.g. Tsai [3], Liu [4] dealt with these heat exchangers with different geometries. The conclusions about temperature and pressure drop were similar to Gherasim [2].

Novosád in work [5] investigated the influence of oblique waves on the heat transfer surface. The biggest problem in this work was the creation of custom geometry. Each option had to be modeled separately and a meshed. Each model had to be loaded into the solver, set the boundary conditions and subsequently evaluated by calculation.

Disadvantages of repeated generation of computational meshes are: It is slow, meshes made in different models are not similar and parameterization of the model is problematic. Further, even a small change of geometry requires to go through the whole process of model creation and mesh generation again. As a result, there is high probability of creation errors of model and low quality of mesh cells. It is necessary to setup the solver, boundary conditions and all models for all computed variants. Furthermore meshes are not similar, i.e. the size, shape, height of wall adjacent cells are not the same for different topologies.

Therefore Dvořák in works [6] and [7] developed a new

Author gratefully acknowledges financial support by Czech Technological Agency under the project TACR TA01020313 and project SGS 21 000.

V. Dvořák is with the Technical university of Liberec, Faculty of mechanical engineering, Department of Power Engineering Equipment. Address: Studentska 2, 46117, Liberec. Phone: +420 485 353 479; e-mail: vaclav.dvorak@tul.cz.

J. Novosád is with the Technical university of Liberec, Faculty of mechanical engineering, Department of Power Engineering Equipment. Address: Studentska 2, 46117, Liberec. E-mail: jan.novosad@tul.cz.

method for generation of computational variants. This method was based on dynamic mesh which is provided by software Fluent. Meshes were created by pulling, which is similar to the own production process, i.e. it is perpendicular to the plates. The main advantages is that such generation of variant is automatic and controlled by in-house software. All computational variants thus have similar mesh.

The aim of this work is to compare numerical results obtained by using two different ways of generation of computational meshes and to find correct mesh parameters to gain accurate numerical results.

II. METHODS

A. Methods of grid generation

In this paper, we discuss the case of a counter flow heat exchanger, which does not have a symmetrical heat transfer area. It is caused by undulations in herringbone pattern. Processes in such heat exchanger can be investigated by modeling the flow around at least two plates using periodical boundary conditions. The first plate has undulations inclined by angle α , while the second one by angle $-\alpha$. How such a model appears can be seen from Fig. 1 and Fig. 2.

In this work we investigated 13 cases of undulations, three different angles of 30° , 45° and 60° and five different pitches of (4, 5, 6, 7 and 8) mm. Definitions of the pitch and angle of undulation are obvious from Fig. 1.

There were 30 waves on each plate. The width of the model B was adjusted for each variant of pitch p and angle α to maintain four crossings of wave peaks of both plates across the model.

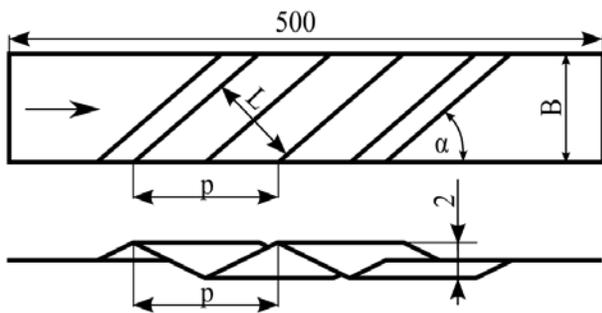


Fig. 1 Plate with oblique ridges - sketch, count of ridges $n = 2$.

The numerical model is in Fig. 2. The heat transfer surface is divided into two parts. Input and output portions (reported as wall) is fixed, and serve to develop the velocity profiles before the central portion (main wall) which will be provided by undulation. Input boundary conditions are specified by velocities (velocity inlets), the output boundary conditions are specified as pressure outlets with static pressure 0 Pa.

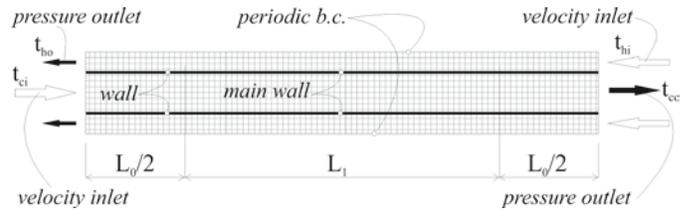


Fig. 2 Model of heat exchange surface of counter flow recuperative heat exchanger.

In this study, we used turbulence model SST $\kappa-\omega$, medium was air considered as incompressible gas. As a results, we obtained pressure, velocity, turbulence and temperature fields inside the computational domain for average inlet velocity of air 2.5 m/s.

Creating computational meshes is based on two different methods. First method is creating the mesh manually. Second method is based on dynamic mesh method provided by software Fluent. The algorithm for mesh creating was described by Dvořák in work [6], [7].

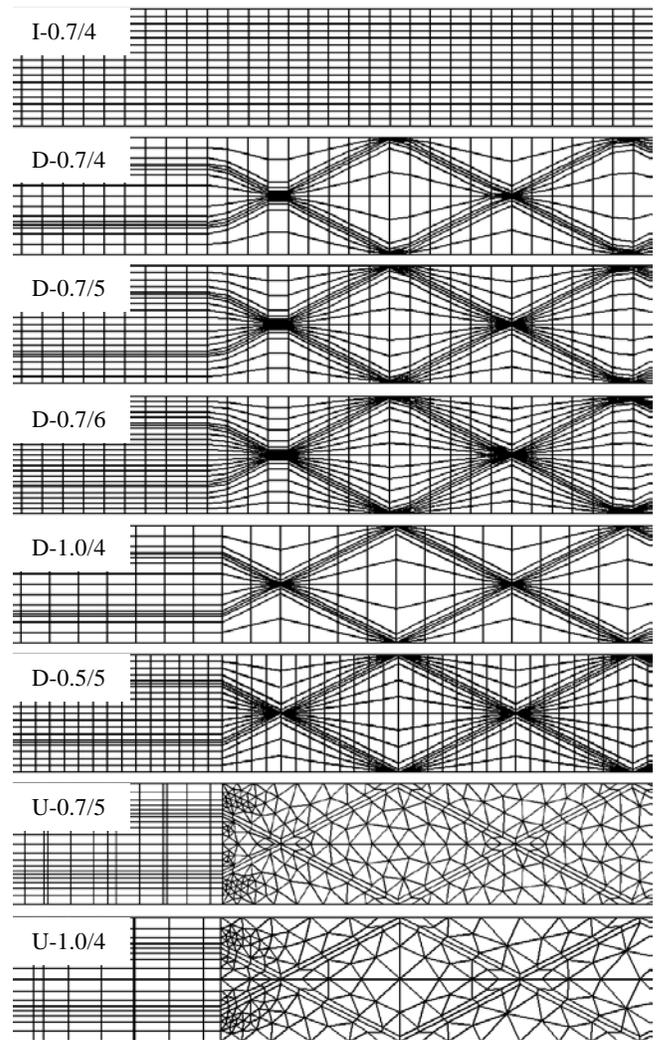


Fig. 3 Computational meshes, from up to bottom – initial mesh before deforming (I-0.7/4), meshes after deforming (D-0.7/4, D-0.7/5, D-0.7/6, D-1.0/4, D-0.5/5), unstructured meshes (U-0.7/5 and U-1.0/4).

Meshes of various density and element size were created. We used very coarse mesh with element size of 1.0 mm with 4 layers of computational cells across the half of the channel, middle fine meshes with element size of 0.7 mm with 4, 5 and 6 layers and very fine mesh with element size of 0.5 mm and 5 layers of computational cells.

We also used two meshes generated manually, a coarse one with element size of 1.0 mm with 4 layer of computational cells and a middle fine mesh with element size of 0.7 mm with 5 layers of computational cells.

To compare obtained numerical results, all meshes had the same width of wall adjacent cells which was 0.1 (mm) and value of y^+ was around 1.5. Comparison of used meshes are in Fig. 3, where cuts of computational meshes are for undulation with angle of 30° and pitch of 4 (mm).

Fig. 4 shows the comparison of the mesh after deforming and unstructured mesh in isometric view, where the meshes are for undulation with angle of 45° , element size of 0.7 (mm), and pitch of 5 (mm).

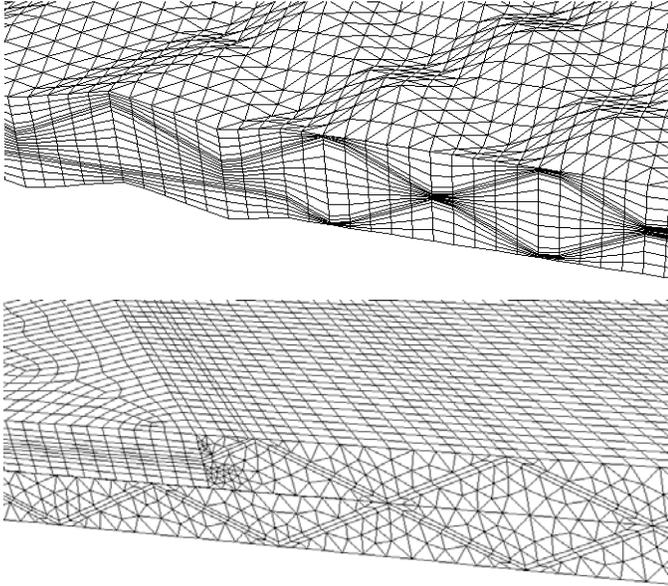


Fig. 4 Computational meshes, from up to bottom –mesh after deforming (D-0.7/5), unstructured meshes (U-0.7/5).

We should mention limitation of both methods. For deformed meshes, it is not possible to ensure that all peaks of undulations are modelled properly, see variants D-0.7/5, D-0.7/6, D-1.0/4 in Fig. 3. Meshes generated manually are generated by pulling meshes, see variants U-0.7/4 and U-1.0/4 in Fig. 2, in direction given by angle α , which means that the quality of the mesh decreases for low angles of undulation.

B. Theory of counter flow heat exchangers

Most of the recuperative heat exchangers in air conditioning works in the isobaric mode, where mass flow rates of warm and cold air are equal, i.e. $\dot{m}_c = \dot{m}_h$. Assuming equality of

specific heat capacities, $c_{pc} = c_{ph}$, we can write the coefficient of efficiency as

$$\eta = \frac{t_{hi} - t_{ho}}{t_{hi} - t_{ci}}, \quad (1)$$

where t_{hi} ($^\circ\text{C}$) is the inlet temperature of hot air. Furthermore index c denotes a cold stream, index i inlet into the heat exchanger and index o the outlet of the heat exchanger.

For the pressure drop assessment, it is used the ratio of total pressure loss between the inlet and outlet

$$\Delta p = \bar{p}_i - \bar{p}_o, \quad (4)$$

where \bar{p}_i (Pa) is mass averaged total pressure in the inlet and \bar{p}_o (Pa) is mass averaged total pressure at the pressure outlet.

The dependence between the heat balance and efficiency η is expressed as

$$k A = \dot{m} c_p \frac{\eta}{1 - \eta} \quad (5)$$

where \dot{m} (kg s^{-1}) is the mass flow rate, c_p ($\text{J kg}^{-1} \text{K}^{-1}$) is isobaric specific heat capacity, k ($\text{W m}^{-2} \text{K}^{-1}$) heat transfer coefficient and A (m^2) is the area of heat transfer surface.

III. RESULTS

Dependency of efficiency on the pitch of undulation for various angles of undulation is shown in Fig. 5, Fig. 6 and Fig. 7. These figures describes the effect of mesh type and elements quality on efficiency value. Mesh types created by deforming (dynamic mesh) are labelled by letter “D”. Unstructured meshes are labelled by letter “U”. Element size and the count of layers of computational cells across the channel are labeled by numbers, e.g. “0.5/5” means that the element is 0.5 mm and there are 5 layers across the channel.

We can see that the efficiency increases for higher pitch of undulations for all angles, because the air flows around the undulations in better way for high pitches.

Generally we can see that difference between all meshes are more significant for lower pitch of undulation.

As we can see there is significant difference between the case “D-1.0/4”, which is the coarsest mesh, and other cases. The efficiency value for this case is higher than for the other ones, so we can assume that this mesh variant is not suitable to use for computations. That could be caused by poor mesh quality, while comparing other results the wall friction is overvalued.

We can observe, by comparing cases for element size 0.7 mm, the effect of count of cell layers. Lower efficiency and so lower heat transfer coefficient is obtained for higher count of computational cells across the channel. Remind that all cases

have the same width of 0.1 mm of adjacent cells and the deformed cases have also the same width of the second cell of 0.3 mm.

case “U-0.7/5”. That could be caused by deformation of elements of unstructured mesh in the direction along the width of heat transfer area.

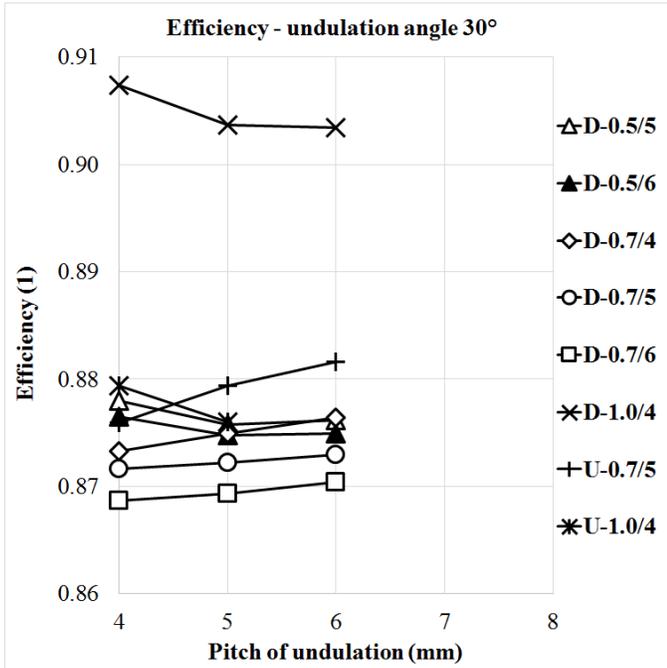


Fig. 5 Dependency of the efficiency of the heat exchanger on pitch of undulations for undulation angle 30°.

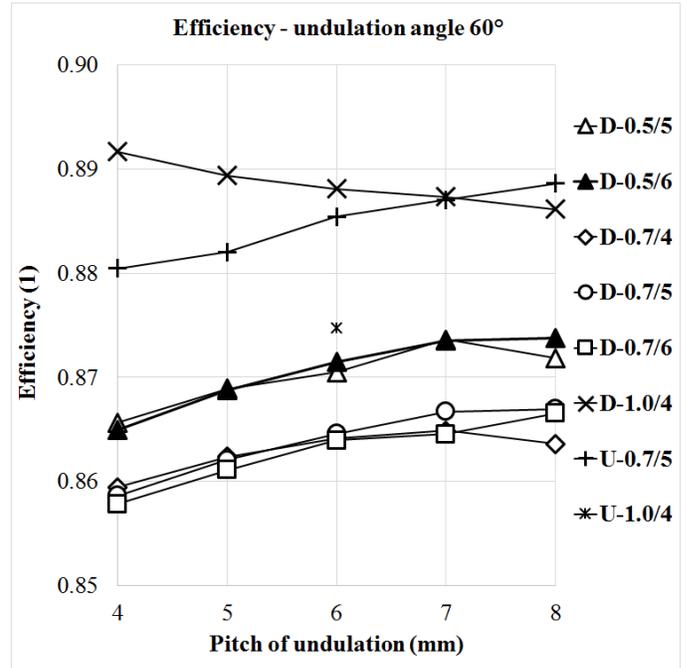


Fig. 7 Dependency of the efficiency of the heat exchanger on pitch of undulations for undulation angle 60°.

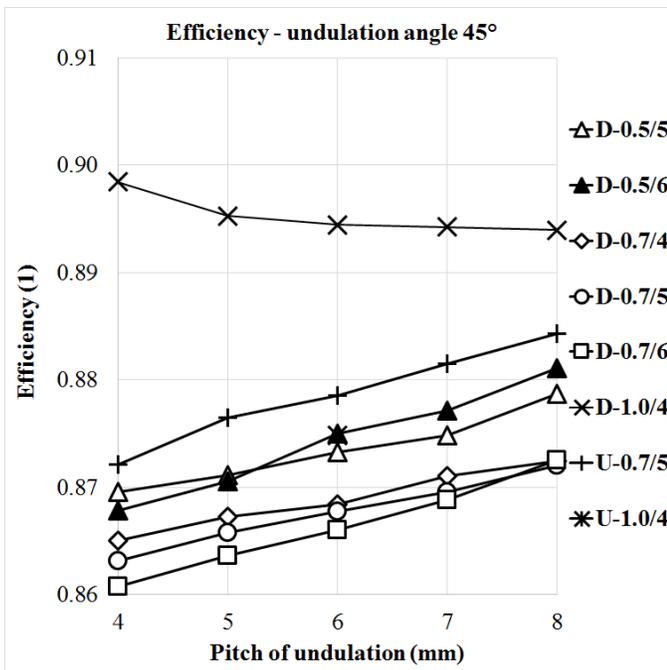


Fig. 6 Dependency of the efficiency of the heat exchanger on pitch of undulations for undulation angle 45°.

An important observation is that the best match of the efficiency values between the mesh creation ways for undulation angle 30° is for case “D-0.5/5” and case “U-0.7/5”.

As we can see from Fig. 6 and Fig. 7, the higher undulation angle causes higher difference between case “D-0.5/5” and

Dependency of pressure drop on the pitch of undulation for various angles of undulation is shown in Fig. 8, Fig. 9 and Fig. 10. These figures describes the effect of mesh type and elements size on pressure drop. Labels of each variant of mesh are done in the same way as for dependency of efficiency.

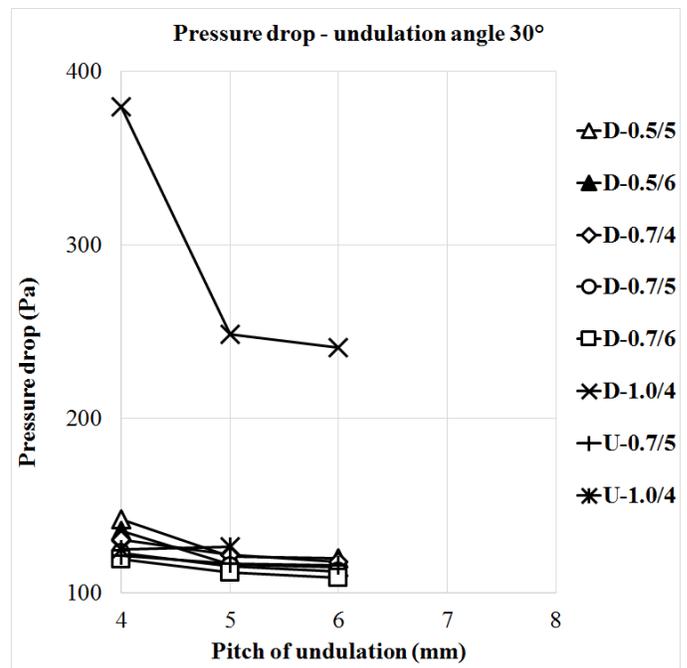


Fig. 8 Dependency of the pressure drop of the heat exchanger on pitch of undulations for undulation angle 30°.

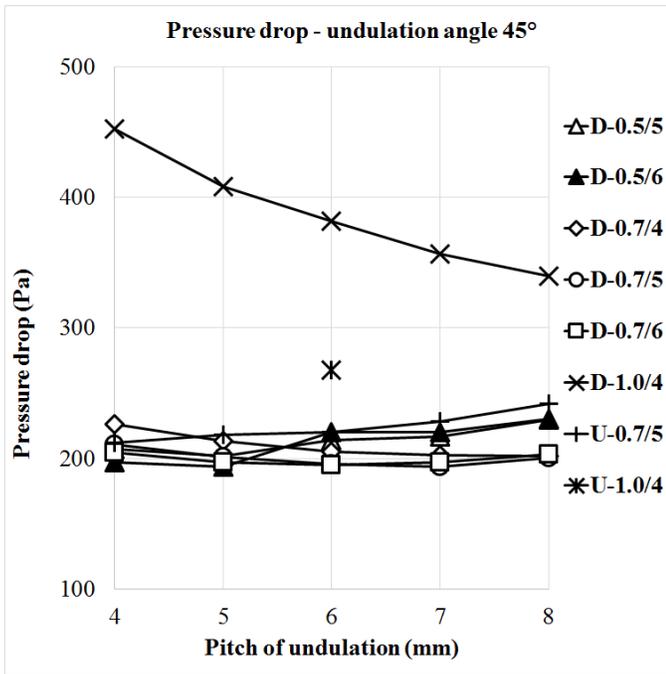


Fig. 9 Dependency of the pressure drop of the heat exchanger on pitch of undulations for undulation angle 45°.

As we can see there is significant difference between the case “D-1.0/4” and other cases for undulation angle 30° and 45°. The pressure drop for this case is higher than for the other ones.

As we can see from Fig. 8 and 9, pressure drop for all cases (except “D-1.0/4”) are in very narrow range. Generally the obtained pressure loss is lower for higher count of layers of computational cells and form finer meshes.

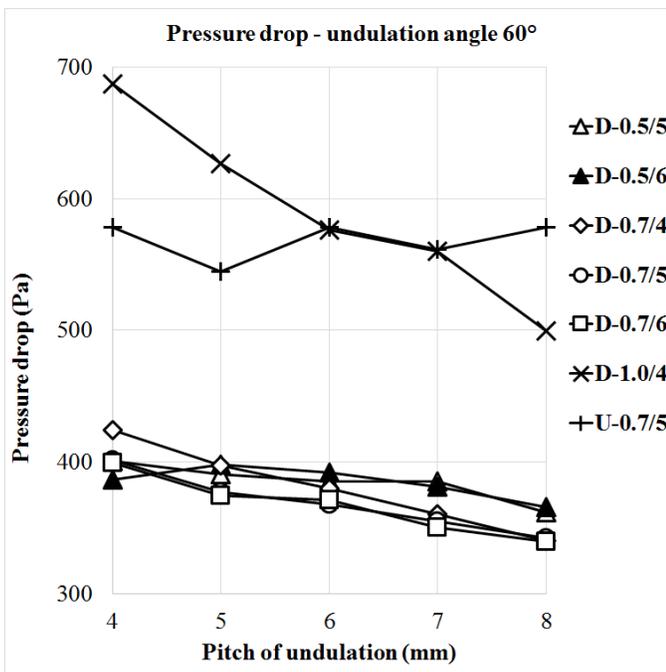


Fig. 10 Dependency of the pressure drop of the heat exchanger on pitch of undulations for undulation angle 60°.

Another situation is shown in Fig. 10. We can see there are two groups of results. A good match between the two coarsest cases “D-1.0/4” and “U-0.7/5”. It is possible that there are some effects which are coupled with the poor mesh quality. These effects caused higher pressure drop, e.g. we evaluated high wall friction compare to other cases.

The second group of results are for finer deformed meshes. As for efficiency, we can see that obtained pressure loss decreases with count of layers of computational cells, but it is higher for the finest mesh D-0.5/5. It seems that differences between variants D-0.7/5 and D-0.7/6 are negligible.

IV. CONCLUSIONS

We investigated flow in counter flow plate heat exchanger. We used CFD and two different method of generation of computational meshes and examined effect of computational cells size and count of layers of computational cells across the half of the channel. For accurate comparison, all meshes had the same width of wall adjacent cells. We used manually generated unstructured meshes and automatically generated dynamic meshes. We investigated flow around plates with various undulations with different angle and pitch of undulation and evaluated pressure loss and efficiency of heat transfer of the exchanger.

Generally the obtained efficiency and obtained pressure loss decreased with higher mesh density, i.e. with lower size of computational cells.

We found that very coarse meshes can yield both too high and unrealistic efficiency and pressure loss compare too other results. It seems that for deformed meshes, which are structured, the satisfactory size of elements is 0.7 mm, while this size can be unsatisfactory for unstructured meshes in some cases.

Comparing meshes with the same density, we can conclude that higher count of layers of computational cells brings lower efficiency and lower pressure loss. While there is still obvious difference in efficiency for counts of layers 5 and 6, it seems that 5 layers are enough to obtain the lowest pressure loss.

The finest structured mesh with element size of 0.5 mm and 5 layers brings slightly higher efficiency and pressure loss than structured meshes with element size 0.7. We can hardly state whether it is because of finer mesh or because of better reproduction of peaks of undulation. Therefore, it could be beneficial to have data also for finest mesh with higher count of layers.

Finally, from trends visible in our results, we can say that meshes with lower efficiency and lower pressure loss seem to be more correct than meshes yielding higher efficiencies and higher pressure losses. This is also confirmed by our general experience and may be explain by more significant effect of numerical dissipation while using coarser meshes. Of course, to confirm this statement, the numerical data should be compared with experiments.

REFERENCES

- [1] T. Vít., P. Novotný, Nguyen Vu, V. Dvořák, "Testing method of materials for enthalpy wheels," *Recent Advances in Energy, Environment, Economics and Technological Innovation*, Paris, France, 29th – 31st October, 2013.
- [2] I. Gherasim, N. Galanis, C. T. Nguyen, "Heat transfer and fluid flow in a plate heat exchanger. Part II: Assessment of laminar and two-equation turbulent models," *International Journal of Thermal Sciences* 50, 2011, p. 1499-1511.
- [3] Y. Ch. Tsai, F. B. Liu, P. T. Shen, "Investigation of effect of oblique ridges on heat transfer in plate heat exchangers," *International Communications in Heat and Mass Transfer* 36, 2009, p. 574–578.
- [4] F. B. Liu, Y. Ch. Tsai, "An experimental and numerical investigation of fluid flow in a cross-corrugated channel," *Heat Mass Transfer* 46, 2010, p. 585–593
- [5] J. Novosád, V. Dvořák, "Investigation of effect of oblique ridges on heat transfer in plate heat exchangers", *Experimental Fluid Mechanics 2013*, November 19.-22., 2013, pp. 510 - 514.
- [6] V. Dvořák, "A method for optimization of plate heat exchanger", *18th International Conference on Circuits, Systems, Communications and Computers*, Santorini Island, Greece, July 17-21, 2014.
- [7] V. Dvořák, "Optimization of plate heat exchangers with angled undulations in herringbone pattern", *The 7th International Meeting on Advances in Thermofluids*, Kuala Lumpur, Malaysia, November, 26-27, 2014.

Sampling Time Dependency of Chaotic Ueda Oscillator as the Generator of Random Numbers for Heuristic

Roman Senkerik, Michal Pluhacek, and Zuzana Kominkova Oplatkova

Abstract— This paper investigates the utilization of the time-continuous chaotic system, which is UEDA oscillator, as the chaotic pseudo random number generator. (CPRNG). Repeated simulations were performed investigating the influence of the UEDA oscillator sampling time to the selected heuristic, which is differential evolution algorithm (DE). Initial experiments were performed on the Schwefel function in higher dimensions.

Keywords— Deterministic chaos; Chaotic oscillators; Heuristic; Differential Evolution; Chaotic Pseudo Random Number Generators

I. INTRODUCTION

GENERALLY speaking, the term “chaos” can denote anything that cannot be predicted deterministically. In the case that the word “chaos” is combined with an attribute such as “deterministic”, then a specific type of chaotic phenomena is involved, having their specific laws, mathematical apparatus and a physical origin. The deterministic chaos is a phenomenon that - as its name suggests - is not based on the presence of a random or any stochastic effects. It is clear from the structure of the equations (see the section 4), that no mathematical term expressing randomness is present. The seeming randomness in deterministic chaos is related to the extreme sensitivity to the initial conditions [1].

Till now, the chaos has been observed in many of various systems (including evolutionary one). Systems exhibiting deterministic chaos include, for instance, weather, biological systems, many electronic circuits (Chua’s circuit), mechanical systems, such as double pendulum, magnetic pendulum, or so called billiard problem.

The idea of using chaotic systems instead of random processes (pseudo-number generators - PRNGs) has been presented in several research fields and in many applications with promising results [2], [3].

R.Senkerik, M. Pluhacek and Z. Kominkova Oplatkova are with the Department of Informatics and Artificial Intelligence, Faculty of Applied Informatics, Tomas Bata University in Zlin, Nam. T.G. Masaryka 5555, 760 01 Zlin, Czech Republic, (phone: +420576035189; fax: +420576035279; e-mail: senkerik@fai.utb.cz).

This work was supported by Grant Agency of the Czech Republic - GACR P103/15/06700S, further by financial support of research project NPU I No. MSMT-7778/2014 by the Ministry of Education of the Czech Republic and also by the European Regional Development Fund under the Project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089.; and by Internal Grant Agency of Tomas Bata University under the project No. IGA/FAI/2015/057.

Another research joining deterministic chaos and pseudorandom number generator has been done for example in [4]. Possibility of generation of random or pseudorandom numbers by use of the ultra weak multidimensional coupling of p 1-dimensional dynamical systems is discussed there. Another paper [5] deeply investigate logistic map as a possible pseudorandom number generator and is compared with contemporary pseudo-random number generators. A comparison of logistic map results is made with conventional methods of generating pseudorandom numbers. The approach used to determine the number, delay, and period of the orbits of the logistic map at varying degrees of precision (3 to 23 bits). Another paper [6] proposed an algorithm of generating pseudorandom number generator, which is called (couple map lattice based on discrete chaotic iteration) and combine the couple map lattice and chaotic iteration. Authors also tested this algorithm in NIST 800-22 statistical test suits and for future utilization in image encryption. In [7] authors exploit interesting properties of chaotic systems to design a random bit generator, called CCCBG, in which two chaotic systems are cross-coupled with each other. A new binary stream-cipher algorithm based on dual one-dimensional chaotic maps is proposed in [8] with statistic proprieties showing that the sequence is of high randomness. Similar studies are also done in [9], [10] and [11].

II. MOTIVATION

Recent research in chaos driven heuristics has been fueled with the predisposition that unlike stochastic approaches, a chaotic approach is able to bypass local optima stagnation. This one clause is of deep importance to evolutionary algorithms. A chaotic approach generally uses the chaotic system in the place of a pseudo random number generator [12]. This causes the heuristic to map unique regions, since the chaotic system iterates to new regions. The task is then to select a very good chaotic system (either discrete or time-continuous) as the pseudo random number generator.

The initial concept of embedding chaotic dynamics into the evolutionary algorithms is given in [13]. Later, the initial study [14] was focused on the simple embedding of chaotic systems in the form of chaos pseudo random number generator (CPRNG) for DE (Differential Evolution) and SOMA [15] in

the task of optimal PID tuning

Several papers have been recently focused on the connection of heuristic and chaotic dynamics either in the form of hybridizing of DE with chaotic searching algorithm [16] or in the form of chaotic mutation factor and dynamically changing weighting and crossover factor in self-adaptive chaos differential evolution (SACDE) [17]. Also the PSO (Particle Swarm Optimization) algorithm with elements of chaos was introduced as CPSO [18] or CPSO combined with chaotic local search [19].

This idea was later extended with the successful experiments with chaos driven DE (ChaosDE) [20], [21] with both and complex simple test functions and in the task of chemical reactor geometry optimization [22].

The concept of Chaos DE has proved itself to be a powerful heuristic also in combinatorial problems domain [23].

At the same time the chaos embedded PSO with inertia weigh strategy was closely investigated [24], followed by the introduction of a PSO strategy driven alternately by two chaotic systems [25] and novel chaotic Multiple Choice PSO strategy (Chaos MC-PSO) [26].

The primary aim of this work is not to develop a new type of pseudo random number generator, which should pass many statistical tests, but to try to test, analyze and compare the implementation of different natural chaotic dynamics as the CPRNGs, thus to analyze and highlight the different influences to the system, which utilizes the selected CPRNG (including the evolutionary computational techniques).

III. CPRNG CONCEPT

The general idea of CPRNG is to replace the default PRNG with the chaotic system. As the chaotic system is a set of equations with a static start position, we created a random start position of the system, in order to have different start position for different experiments. This random position is initialized with the default PRNG, as a one-off randomizer. Once the start position of the chaotic system has been obtained, the system generates the next sequence using its current position.

Generally there exist many other approaches as to how to deal with the negative numbers as well as with the scaling of the wide range of the numbers given by the chaotic systems into the typical range 0 – 1:

- Finding of the maximum value of the pre-generated long discrete sequence and dividing of all the values in the sequence with such a maxval number.
- Shifting of all values to the positive numbers (avoiding of ABS command) and scaling.

IV. UEDA OSCILLATOR

UEDA oscillator is the simple example of driven pendulums, which represent some of the most significant examples of chaos and regularity.

The UEDA system can be simply considered as a special

case of intensively studied Duffing oscillator that has both a linear and cubic restoring force. Ueda oscillator represents the both biologically and physically important dynamical model exhibiting chaotic motion. It can be used to explore much physical behavior in biological systems. [27]

The UEDA chaotic system equations are given in (1). The parameters are: $a = 1.0$ $b = 0.05$, $c = 7.5$ and $\omega = 1.0$ as suggested in [28]. The x, y parametric plot of the chaotic system is depicted in Fig. 1.

$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= -ax^3 - by + c \sin \omega t \end{aligned} \quad (2)$$

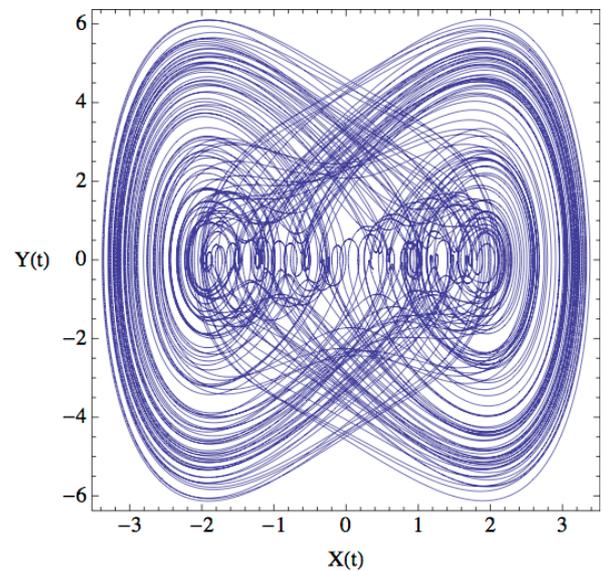


Fig. 1 x, y parametric plot of the UEDA oscillator

V. USED HEURISTIC - DIFFERENTIAL EVOLUTION

This section briefly describes the used heuristic within all experiments.

DE is a population-based optimization method that works on real-number-coded individuals [29]. For each individual $\vec{x}_{i,G}$ in the current generation G , DE generates a new trial individual $\vec{x}'_{i,G}$ by adding the weighted difference between two randomly selected individuals $\vec{x}_{r1,G}$ and $\vec{x}_{r2,G}$ to a randomly selected third individual $\vec{x}_{r3,G}$. The resulting individual $\vec{x}'_{i,G}$ is crossed-over with the original individual $\vec{x}_{i,G}$. The fitness of the resulting individual, referred to as a perturbed vector $\vec{u}_{i,G+1}$, is then compared with the fitness of $\vec{x}_{i,G}$. If the fitness of $\vec{u}_{i,G+1}$ is greater than the fitness of $\vec{x}_{i,G}$, then $\vec{x}_{i,G}$ is replaced with $\vec{u}_{i,G+1}$; otherwise, $\vec{x}_{i,G}$

remains in the population as $\vec{x}_{i,G+1}$. DE is quite robust, fast, and effective, with global optimization ability. It does not require the objective function to be differentiable, and it works well even with noisy and time-dependent objective functions. Please refer to [29], [30] for the detailed description of the used DERand1Bin strategy (2) (both for Chaos DE and Canonical DE) as well as for the complete description of all other strategies.

$$u_{i,G+1} = x_{r1,G} + F \cdot (x_{r2,G} - x_{r3,G}) \quad (2)$$

VI. EXPERIMENT DESIGN

For the purpose of evolutionary algorithm performance comparison within this initial research, the multimodal Schwefel's test function (3) was selected.

$$f(x) = \sum_{i=1}^D -x_i \sin(\sqrt{|x_i|}) \quad (3)$$

Function minimum: Position for E_n :

$$(x_1, x_2, \dots, x_n) = (420.969, 420.969, \dots, 420.969)$$

$$\text{Value for } E_n: y = -418.983 \cdot \text{Dimension}$$

In this paper, the canonical DE strategy DERand1Bin and the Chaos DERand1Bin (ChaosDE) strategy driven by discretized UEDA oscillators (ChaosDE) were used.

The parameter settings for both canonical DE and ChaosDE were obtained analytically based on numerous experiments and simulations (see Table 1). Experiments were performed in the environment of *Wolfram Mathematica*; canonical DE therefore has used the built-in *Wolfram Mathematica* pseudo random number generator *Wolfram Cellular Automata* representing traditional pseudorandom number generator in comparisons. All experiments used different initialization, i.e.

different initial population was generated within the each run of Canonical or Chaos driven DE. The maximum number of generations was fixed at 3000 generations. This allowed the possibility to analyze the progress of DE within a limited number of generations and cost function evaluations.

TABLE I. DE SETTINGS

DE Parameter	Value
Dimension	10
PopSize	50
F	0.8
CR	0.8
Generations	1000
Max. CF Evaluations (CFE)	50000

VII. EXPERIMENT RESULTS

The statistical results of the experiments are shown in Table 2, which represent the simple statistics for cost function (CF) values, e.g. average, median, maximum values, standard deviations and minimum values representing the best individual solution for all 50 repeated runs of canonical DE and several versions of ChaosDE (with several sampling times of UEDA oscillator).

Table 3 compares the progress of several versions of ChaosDE, and Canonical DE. This table contains the average CF values for the generation No. 250, 500, 750 and 1000 from all 50 runs. The bold values within the both Tables 2 and 3 depict the best obtained results. Following versions of Multi-ChaosDE were studied:

Figures 2 a) – 2f) show the influence of sampling rate to the distribution of numbers given by particular CPRNG.

Finally the graphical comparison of the time evolution of average CF values for all 50 runs of ChaosDE and canonical DERand1Bin strategy is depicted in Fig. 3.

TABLE II. SIMPLE RESULTS STATISTICS FOR THE SCHWEFEL'S FUNCTION – 10D

DE Version	Avg CF	Median CF	Max CF	Min CF	StdDev
Canonical DE	-4189.62	-4189.82	-4188.7	-4189.83	0.425073
ChaosDE - Sampling 0.1s	-4098.64	-4189.83	-3559.49	-4189.83	209.0547
ChaosDE - Sampling 0.5s	-3766.93	-3710.45	-3307.99	-4189.67	377.1445
ChaosDE - Sampling 1.0s	-3480.44	-3411.2	-2881.27	-4189.69	436.9247
ChaosDE - Sampling 2.0s	-3587.11	-3554.31	-3092.63	-4187.08	315.9573

TABLE III. COMPARISON OF PROGRESS TOWARDS THE MINIMUM FOR THE SCHWEFEL'S FUNCTION

DE Version	Generation No.: 250	Generation No.: 500	Generation No.: 750	Generation No.: 1000
Canonical DE	-3024.34	-3502.34	-4017.97	-4189.62
ChaosDE - Sampling 0.1s	-2879.33	-3574.31	-3979.95	-4098.64
ChaosDE - Sampling 0.5s	-2609.5	-2909.7	-3366.64	-3766.93
ChaosDE - Sampling 1.0s	-2574.59	-2800.19	-3132.67	-3480.44
ChaosDE - Sampling 2.0s	-2615.01	-2926.31	-3223.74	-3587.11

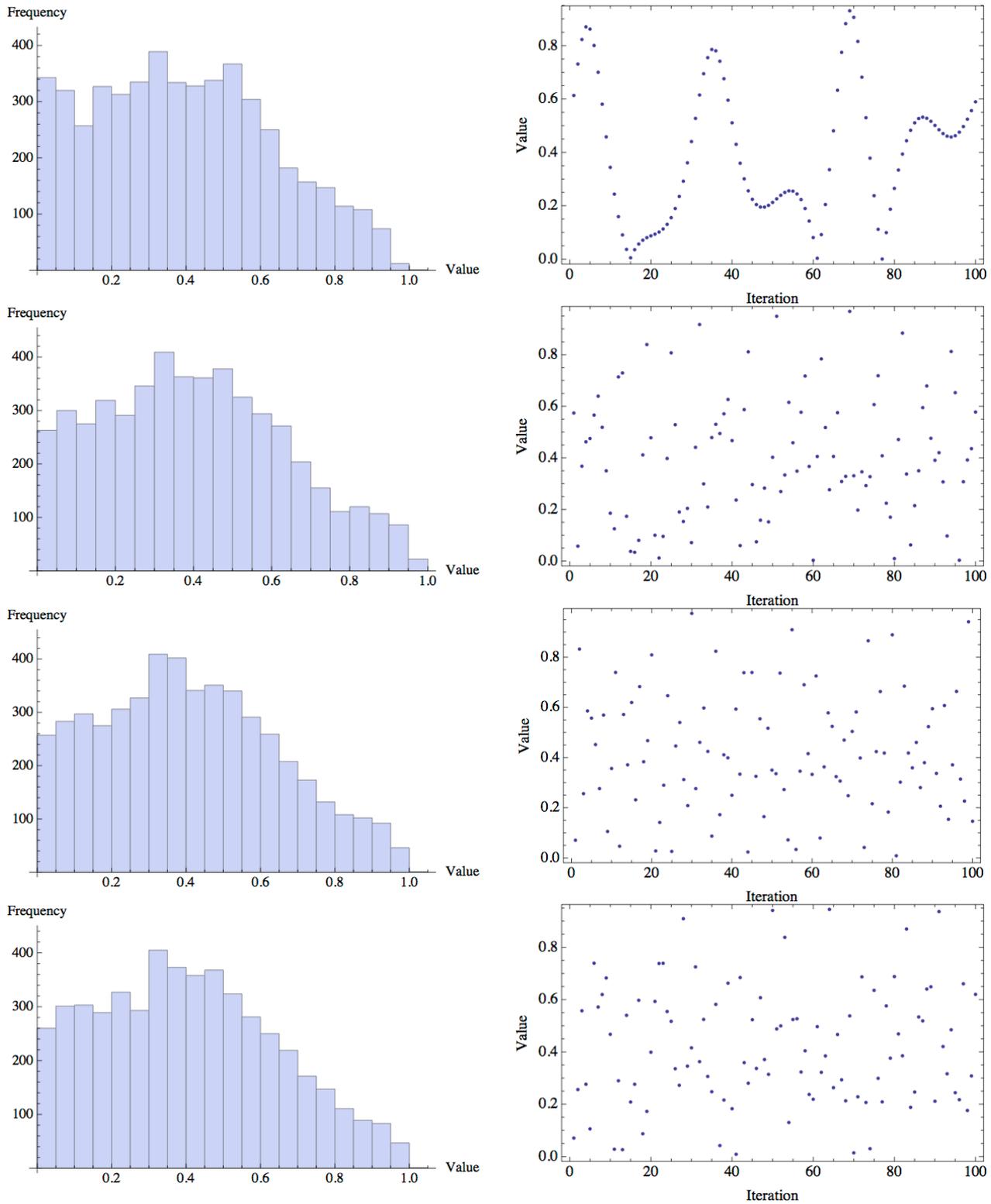


Fig. 2. Comparison of the influence of sampling rate to the distribution of numbers given by UEDA CPRNG; Left: Histogram of the distribution of real numbers transferred into the range $<0 - 1>$; Right: Example of the chaotic dynamics: range $<0 - 1>$ generated by means of UEDA oscillator sampled with the particular sampling rate – variable x ; Sampling rates from up to down: 0.1s, 0.5s, 1.0s, 2.0s.

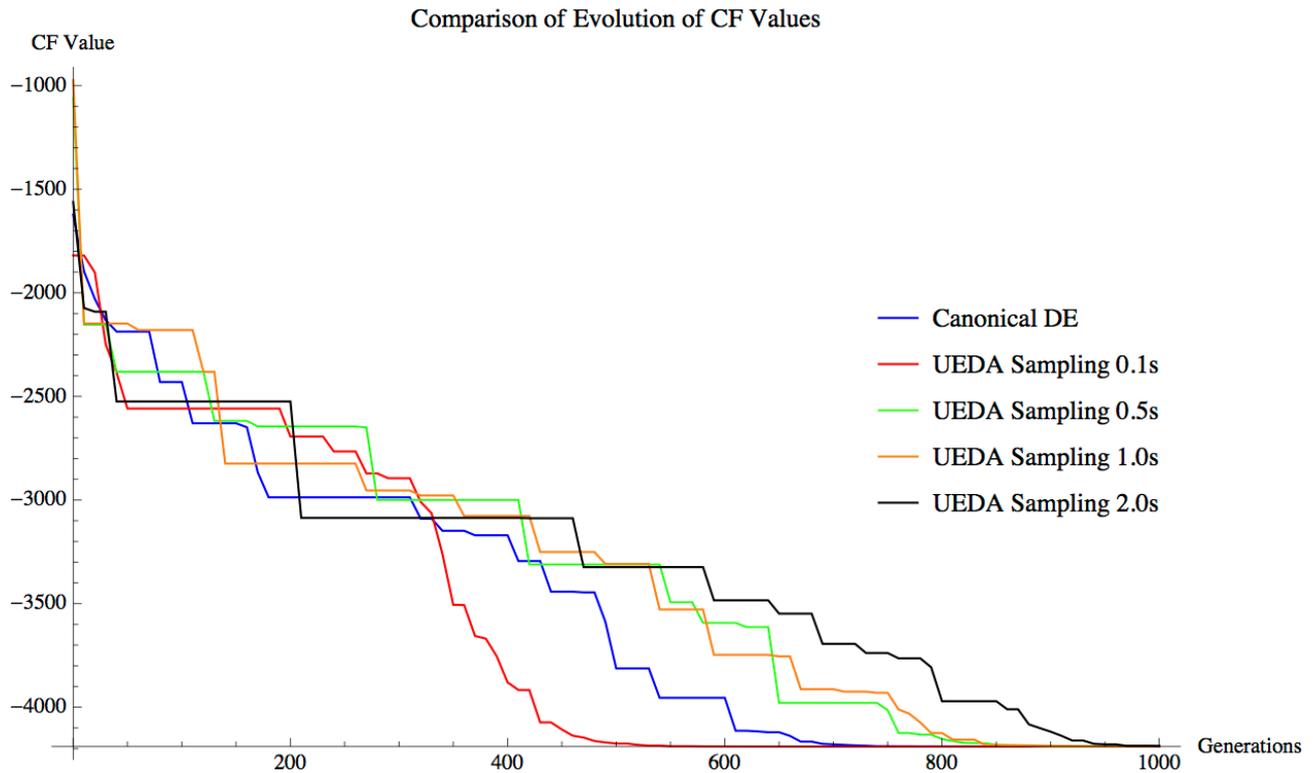


Fig. 3. Comparison of the time evolution of avg. CF values for the all 50 runs of Canonical DE and several versions of ChaosDE driven by UEDA oscillators with different settings of sampling time; Schwefel's function, $Dimension = 10$.

VIII. CONCLUSION

The novelty of this research represents investigating the influence of the chaotic oscillator sampling time used as the CPRNG to the selected heuristic, which is DE.

In this paper, the concept of chaos driven DERand1Bin strategy was more experimentally analyzed and compared with the canonical DERand1Bin strategy on the selected benchmark function with four different settings of sampling time for the UEDA chaotic oscillator.

Obtained numerical results given in Tables 2 and 3 and graphical comparisons in Figures 2 and 3 support the claim that chaos driven heuristic is very sensitive to the hidden chaotic dynamics driving the CPRNG. Such a chaotic dynamics can be significantly changed by the selection of sampling time in the case of the time-continuous systems.

Future plans are including the testing of combination of different time-continuous chaotic systems as well as the adaptive switching and obtaining a large number of results to perform statistical tests.

Furthermore chaotic systems have additional parameters, which can be tuned. This issue opens up the possibility of examining the impact of these parameters to generation of random numbers, and thus influence on the results obtained by means of ChaosDE.

REFERENCES

- [1] S. Celikovsky and I. Zelinka, "Chaos Theory for Evolutionary Algorithms Researchers," in *Evolutionary Algorithms and Chaotic Systems*. vol. 267, I. Zelinka, S. Celikovsky, H. Richter, and G. Chen, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 89-143.
- [2] J. S. Lee and K. S. Chang, "Applications of chaos and fractals in process systems engineering," *Journal of Process Control*, vol. 6, pp. 71-87, 1996.
- [3] J. Wu, J. Lu, and J. Wang, "Application of chaos and fractal models to water quality time series prediction," *Environmental Modelling & Software*, vol. 24, pp. 632-636, 2009.
- [4] R. Lozi, "Emergence of Randomness from Chaos," *International Journal of Bifurcation and Chaos*, vol. 22, p. 1250021, 2012.
- [5] K. J. Persohn and R. J. Povinelli, "Analyzing logistic map pseudorandom number generators for periodicity induced by finite precision floating-point representation," *Chaos, Solitons & Fractals*, vol. 45, pp. 238-245, 2012.
- [6] X.-y. Wang and X. Qin, "A new pseudo-random number generator based on CML and chaotic iteration," *Nonlinear Dynamics*, vol. 70, pp. 1589-1592, 2012/10/01 2012.
- [7] K. P. Narendra, P. Vinod, and K. S. Krishan, "A Random Bit Generator Using Chaotic Maps," *International Journal of Network Security*, vol. 10, pp. 32 - 38, 2010.
- [8] L. Yang and X.-Y. Wang, "Design of Pseudo-random Bit Generator Based on Chaotic Maps," *International Journal of Modern Physics B*, vol. 26, p. 1250208, 2012.
- [9] M. Bucolo, R. Caponetto, L. Fortuna, M. Frasca, and A. Rizzo, "Does chaos work better than noise?," *Circuits and Systems Magazine, IEEE*, vol. 2, pp. 4-19, 2002.

- [10] H. Hu, L. Liu, and N. Ding, "Pseudorandom sequence generator based on the Chen chaotic system," *Computer Physics Communications*, vol. 184, pp. 765-768, 2013.
- [11] A. Pluchino, A. Rapisarda, and C. Tsallis, "Noise, synchrony, and correlations at the edge of chaos," *Physical Review E*, vol. 87, p. 022910, 2013.
- [12] I. Aydin, M. Karakose, and E. Akin, "Chaotic-based hybrid negative selection algorithm and its applications in fault and anomaly detection," *Expert Systems with Applications*, vol. 37, pp. 5285-5294, 2010.
- [13] R. Caponetto, L. Fortuna, S. Fazzino, and M. G. Xibilia, "Chaotic sequences to improve the performance of evolutionary algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 7, pp. 289-304, 2003.
- [14] D. Davendra, I. Zelinka, and R. Senkerik, "Chaos driven evolutionary algorithms for the task of PID control," *Computers & Mathematics with Applications*, vol. 60, pp. 1088-1104, 2010.
- [15] I. Zelinka, "SOMA — Self-Organizing Migrating Algorithm," in *New Optimization Techniques in Engineering*. vol. 141, ed: Springer Berlin Heidelberg, 2004, pp. 167-217.
- [16] W. Liang, L. Zhang, and M. Wang, "The chaos differential evolution optimization algorithm and its application to support vector regression machine," *Journal of Software*, vol. 6, pp. 1297- 1304, 2011.
- [17] G. Zhenyu, C. Bo, Y. Min, and C. Binggang, "Self-Adaptive Chaos Differential Evolution," in *Advances in Natural Computation*. vol. 4221, L. Jiao, L. Wang, X.-b. Gao, J. Liu, and F. Wu, Eds., ed: Springer Berlin Heidelberg, 2006, pp. 972-975.
- [18] L. d. S. Coelho and V. C. Mariani, "A novel chaotic particle swarm optimization approach using Hénon map and implicit filtering local search for economic load dispatch," *Chaos, Solitons & Fractals*, vol. 39, pp. 510-518, 2009.
- [19] W.-C. Hong, "Chaotic particle swarm optimization algorithm in a support vector regression electric load forecasting model," *Energy Conversion and Management*, vol. 50, pp. 105-117, 2009.
- [20] R. Senkerik, M. Pluhacek, I. Zelinka, Z. Oplatkova, R. Vala, and R. Jasek, "Performance of Chaos Driven Differential Evolution on Shifted Benchmark Functions Set," in *Proc. International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*. vol. 239, Á. Herrero, B. Baruque, F. Klett, A. Abraham, V. Snášel, A. C. P. L. F. Carvalho, *et al.*, Eds., ed: Springer International Publishing, 2014, pp. 41-50.
- [21] R. Senkerik, D. Davendra, I. Zelinka, M. Pluhacek, and Z. Kominkova Oplatkova, "On the Differential Evolution Driven by Selected Discrete Chaotic Systems: Extended Study," in *Proc. 19th International Conference on Soft Computing, MENDEL 2013*, 2013, pp. 137-144.
- [22] R. Senkerik, M. Pluhacek, Z. K. Oplatkova, D. Davendra, and I. Zelinka, "Investigation on the Differential Evolution driven by selected six chaotic systems in the task of reactor geometry optimization," in *Proc. 2013 IEEE Congress on Evolutionary Computation (CEC)*, 2013, pp. 3087-3094.
- [23] D. Davendra, M. Bialic-Davendra, and R. Senkerik, "Scheduling the Lot-Streaming Flowshop scheduling problem with setup time with the chaos-induced Enhanced Differential Evolution," in *Proc. 2013 IEEE Symposium on Differential Evolution (SDE)*, 2013, pp. 119-126.
- [24] M. Pluhacek, R. Senkerik, D. Davendra, Z. Kominkova Oplatkova, and I. Zelinka, "On the behavior and performance of chaos driven PSO algorithm with inertia weight," *Computers & Mathematics with Applications*, vol. 66, pp. 122-134, 2013.
- [25] M. Pluhacek, R. Senkerik, I. Zelinka, and D. Davendra, "Chaos PSO algorithm driven alternately by two different chaotic maps - An initial study," in *Proc. 2013 IEEE Congress on Evolutionary Computation (CEC)*, 2013, pp. 2444-2449.
- [26] M. Pluhacek, R. Senkerik, and I. Zelinka, "Multiple Choice Strategy Based PSO Algorithm with Chaotic Decision Making – A Preliminary Study," in *Proc. International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*. vol. 239, Á. Herrero, B. Baruque, F. Klett, A. Abraham, V. Snášel, A. C. P. L. F. Carvalho, *et al.*, Eds., ed: Springer International Publishing, 2014, pp. 21-30.
- [27] L. Bharti and M. Yuasa. Energy Variability and Chaos in Ueda Oscillator.
Available: <http://www.rist.kindai.ac.jp/no.23/yuasa-EVCUO.pdf>
- [28] J. C. Sprott, *Chaos and Time-Series Analysis*: Oxford University Press, 2003.

The application of Business Intelligence systems in the support of decision processes in the international enterprises

Leszek Ziara

Abstract—The aim of this paper is to focus on the application of Business Intelligence systems in the support of decision making process in the international enterprises. It provides brief characteristics of BI systems, advantages resulting from its application in the management of international enterprises and review of foreign research and case studies regarding the subject. It also provides empirical research concerning the role of BI systems in supporting decision making process in the international enterprises.

Keywords — Business Intelligence systems, business analytics, big data, data warehousing, data mining, decision support systems.

I. INTRODUCTION

NOWADAYS many international enterprises have implemented Business Intelligence systems of different vendors in order to achieve multiple advantages such as improvement of decision making processes at strategic, tactical and operational level of management, gaining competitive advantage on the local and international markets, improvement of communication efficiency and efficacy among different branches of particular enterprise and so on. The most important feature of BI systems in the support of decision making process is acceleration of this process at all levels of management. Faster making of decisions means e.g. the possibility of lowering the cost of enterprise's functionality. T. Davenport underlines significance of information which is indispensable in the decision making and claims that the aim of BI and other decision support systems is "that better information would lead to better decisions and better ways of managing organizational processes" [1] and he further mentions that "if the goal of better information and better analysis is ultimately better decisions and actions taken based on them, organizations must have a strong focus on decisions and their linkage to information" [1]. R. Skyrius et. al. underline relationship between Management Decision Support and Business Intelligence and state that "management decision making is an information-intensive activity, where the structuredness of the problem to be solved directly translates to the complexity of the information tasks to produce a well

supported decision"[2].

II. CHARACTERISTIC AND ADVANTAGES OF BI SYSTEMS APPLICATION IN THE DECISION MAKING SUPPORT

There can be found many definitions of Business Intelligence. Most often it is perceived as provision of the right information to the right people at the right time. The author of the term Business Intelligence is Howard Dresner from Gartner Group who introduced it in 1989. E. Turban perceives BI as "a broad category of applications and techniques for gathering, storing, analyzing and providing access to data to help enterprise user make better business and strategic decisions"[3]. S. Rouhani et. al claim that "Business Intelligence is a managerial concept which refers to a set of programs and technologies that provide capabilities of gathering, analyzing and accessing data of organization's processes (...) BI helps to organizations which having comprehensive knowledge about business affecting factors, such as standards in selling, production and internal organization's processes. (...) The ultimate goal of business intelligence systems in any organization is to help making optimal decisions as soon as possible and in all organization's levels" [4].

The fundamental advantages related to the application of Business Intelligence systems in the management of enterprises may concern: "getting in one place reliable and coherent data and information from all areas of organization's activity which is connected with aspects of systems integration, facilitated access to data coming from different dispersed sources, shortening the time of different analyses, decision making and increasing efficiency and efficacy of management, efficient planning, simulations and prognoses in different angles, instant reaction to appearing market trends, detection of threats and chances in the area of leading activity, current analysis of financial situation and tracking budget deviations, financial optimization of undertaken activities, lowering the number of persons involved in decision making processes, influence on income growth, reduction of costs and improvement of customer's satisfaction" [5]. Business Intelligence systems allow for efficient data transformation into valuable information and as its consequence enable acquisition of knowledge indispensable for making efficient decision. As a result BI systems constitute a solution allowing

Author is with the Faculty of Management of Czestochowa University of Technology, Czestochowa, Poland (e-mail: ziora@zim.pcz.pl).

organizations for using the potential contained in information resources and for allowing employees to acquire up to date knowledge on enterprise condition as well as its market environment [6]. Business Intelligence systems can also be useful in creation, modification and improvement of enterprises' strategy as well as in the management of key business processes and its optimization with the usage of e.g. real time BI systems, semantic BI, hybrid systems and big data solutions.

Business Intelligence may be perceived as a competitor differentiator. Such systems as it was mentioned can contribute to achievement of competitive advantage of a company which have implemented it. E. Turban, R. Sharda, J. Aronson and D. King put emphasis on e.g. strategic imperative of BI and claim that BI have significant value to organizations and "in addition to ROI and other tangible benefits there is increasing evidence that BI are becoming a strategic imperative" [7, p.22-23]. The authors underline importance of competitor analysis which constitute a base for strategic planning. Such solutions can help to sustain competitive advantage in different industries as well. It is worth mentioning the success factors of BI system implementation. E. Turban et al. mention that "the success of BI depends, in part, on which personnel in the organization would be the most likely to make use of it and it must be of benefit to the enterprise as a whole" [7, p.24] C. Olszak and E. Ziemia draw conclusions on the basis of research on the sample of small and medium sized enterprises that "'quick decision-making in enterprise gives a chance to overtake the competition, and it is possible when the managers have free access to business information, and it is the result of analysis of massive amounts of data; such analyses are well performed by BI systems". The other conclusions drew by the authors is that "small and medium enterprises need BI as well as large enterprises and in order to maintain a competitive advantage it became a must to implement BI system in the enterprises. To keep offer of a company competitive they must among others take decisions quickly [8, p.141]. N. Yogev, A. Even, L. Fink state that Business Intelligence systems create value for organizations of different industries. They claim that such systems "represent the natural evolution of decision support systems (DSS) and put a strong emphasis on data-driven decision making, based on the integration of multiple data resources that reflect different aspects of organizational activity" [9]. The authors further state that "BI is unique in its potential to generate both strategic and operational value through the seamless integration of organizational data to support decisions at different levels" [9].

As far as architecture and its role in the decision support is concerned it is worth outlining the exemplary architecture of such solution. In presented scheme (Fig. 1) the data from all the transactional systems operated in a company and in case of international enterprises from its branches as well undergo the process of ETL - extraction, transformation and load, then data is directed into corporate data warehouse and as a final part on the basis of stored data different analyses with the help of data

mining methods and techniques are performed, reports concerning business activity are created, visualization tools, managerial dashboards improve communications within the enterprise. The whole environment is being used to make decision at strategic, tactical and operational level of management. At every stage BI systems support the process of making decision which may be presented as identification of decision problem, finding alternative solutions and finally making a choice.

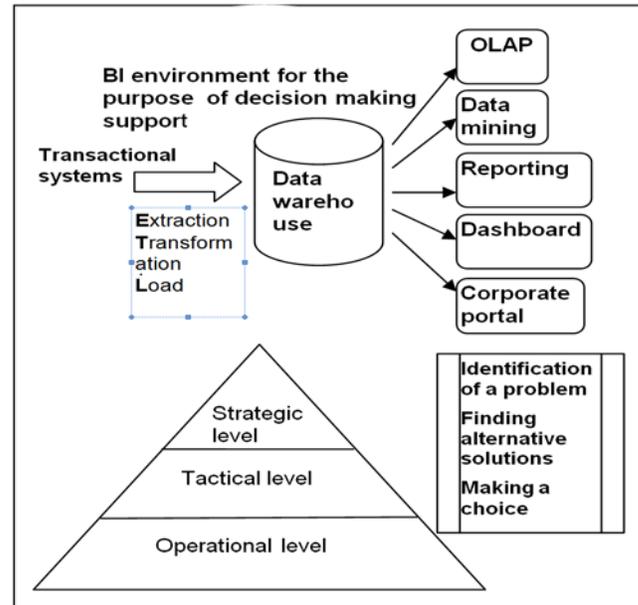


Fig. 1 Example of BI environment for the purpose of decision making

Source: Author's own study

The decision making process in international enterprises embraces the headquarter of enterprises and its branches located in different countries and undergoing different law regulations. In the international companies BI systems should take into consideration different organizational structures and the types of decisions resulting from it. The crucial is the fact that the organizational structure is the main factor of strategy creation [10].

The Business Intelligence sector report by Redwood Capital company states that "over the last few years Business Intelligence has become the top ranked IT priority among enterprise business leaders" [10]. The reports announces that "the global business intelligence market is projected to reach \$20.81 billion in 2018, up from \$13.98 billion in 2013, representing a Compound Annual Growth Rate (CAGR) of 8.28%2. Among all regions, North America is the largest, capturing 49% of the global BI market. It presents the fact that the BI market is segmented into traditional, mobile, cloud and social Business Intelligence, based on product architecture and user interface" [11].

III. REVIEW OF FOREIGN RESEARCH AND CASE STUDIES

On the basis of selected foreign case studies involving the application of BI it is worth mentioning the solution presented by Enterprise Iron company and concerning the application of

BI for large international chemical company. The vendor claims that "for a large chemical company based internationally, a solution for data warehousing and business intelligence was needed to standardize across the enterprise. It was decided that they needed an enterprise solution for these capabilities to provide the competitive advantage gaining benefits of effective business intelligence. (...) EI's Business Intelligence Solutions practice helped the client write an RFP and evaluate potential vendors, ultimately selecting SAP Business Warehouse as the COTS solution. EI's Data Warehousing practices helped establish the plans for migrating data to SAP BW, and were involved in the specification of solutions to cleanse the data and improve customer data quality. EI's Program Management practice provided steering and program oversight to the project through its implementation. (...) As a result of engaging EI, the client consolidated its disparate business intelligence offerings into a standardized industry leading platform, and was able to provide business intelligence on a level not previously possible, exposing insights that promised to enable greater competitive advantage in their market place, while enabling them to diagnose and understand their own operations and customer relationship management for opportunities to improve and modernize sales and operations"[12].

G. Miller et al. presented the results of the survey conducted by BetterManagement subscribers who completed a survey concerning application of BI in their organization. The authors of state that "the online survey was completed by 220 companies across various industries, sizes, geographic locations, and job levels and 84% companies have implemented BI system" [13]. They further claim that "among those companies that conduct a formal BI needs analysis, three quarters (73%) perform an ongoing review of BI needs to ensure that new opportunities or requirements are identified and added to BI processes. (...)The primary components of a BI needs analysis are reports (73%) and strategic analytics (68%). Somewhat less frequently used in the analysis are compliance/corporate governance issues (46%) and early warning systems (44%). Fewer than one in three companies (29%) incorporate legal/ regulatory reporting. The survey also showed that "BI usage still seems to be restricted to management level in most companies" [13].

The other research conducted by W. Eckerson and entitled Business-Driven BI "examines best practices for implementing self-service BI and the technologies and tools that let users create their own reports and dashboards and conduct their own analyses. The survey was taken by 249 people. Survey results are based on 234 respondents who indicated their positions as "BI or IT professional," "BI sponsor or user" or "BI consultant." Responses from those who selected "BI vendor" or "Other" were excluded from the results [14]. (...) The industry with the highest percentage of respondents was manufacturing with 13%. Next were consulting with 11%, retail with 9% and banking, health care and software, all at 8%. Self-Service BI allows users to create own reports and dashboards so they get the information they want, when they want it and how they want it displayed. Self-service BI removes IT professionals as intermediaries between business

users and the data. This gives business users direct access to the raw material" [14]. (...) "There are two types of self-service BI, one for report users and another for report authors. The number one challenge cited by almost three-quarters, or 73%, of BI professionals is counterintuitive: Self-service BI "requires more training than expected. Survey respondents also said self-service BI "creates report chaos" (61%), "makes it harder to find the right report" (36%) and the "tools confuse users" (42%). The research showed that there is a correlation between success with self-service BI and BI adoption rates" [14]. The author drew the conclusion that "self-service BI can empower users and increase BI adoption, but it is difficult to implement properly because there are many types of users with different information requirements. There is no single tool or approach to self-service BI that works in all situations" [14].

Another example of BI application in the organizations is an example of healthcare industry. P. Dindigal presents BI application in saving people's lives and claims that "time is perhaps the most important factor when people are stricken with heart attacks" [15]. He says that "thanks to computer systems that enable clinicians to analyze treatment procedures for suspected coronary victims, Florida's BayCare Hospital has reduced the time it takes to diagnose and process a heart patient by 20 minutes [15]". He further states that "reducing medication errors is another priority for healthcare providers. In Cincinnati, the Children's Hospital Medical Center cut medication errors in half by analyzing orders and feeding the results back to a medication-administration system [15]". "In Tulsa, OK, the St. John Medical Center recently reduced the number of transfusions leading to negative reactions by 18 percent. That saves \$1.4 million annually, and eliminates the opportunity for numerous errors" [15]. The author indicates benefits of BI systems in healthcare such as "faster data gathering and meaningful analytical report production helps in decision support and operational management, while seamless integration and pre-data integration efforts cleanse data and remove duplicate data from various sources. They also provide high-quality data for enterprise decision making. Further, performance and quality improve by segregating main subject areas into key performance indicators (KPIs)" [15]. It should be remembered that "Information Systems of hospital must possess appropriate degree of technological advancement which allows for gathering and delivery of useful data which may be used by BI systems" [16].

IV. EMPIRICAL RESEARCH

In order to prove that Business Intelligence systems have a significant influence on acceleration of decision making process at strategic, tactical and operational level of management there was conducted a survey research on the sample of 34 international enterprises which embraced different industries such as e.g. energetics, electronics, building and construction, retail sale, banking, clothing, IT, logistics, metallurgy, chemical etc. The research was conducted for the purpose of PhD thesis realization of the author. Most of surveyed enterprises belonged to the group of large enterprises. The results showed that thanks to the implementation of Business Intelligence system in a particular

international enterprise the decision making process at all level of management was accelerated. The average values for 3 levels of management is presented in Fig. 2

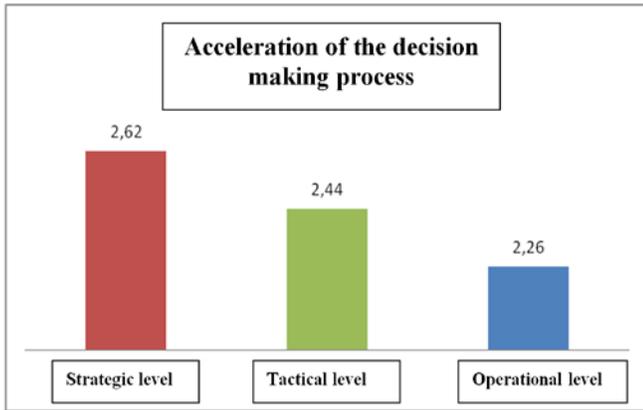


Fig. 2 The average level of decision making support by BI systems with the division into three levels of management

Source: Author's own study

As far as the methodology is concerned the respondents had to select in the research questionnaire whether the decision making process in a given enterprise was greatly, moderately not significantly, was not at all accelerated or it was slowed down after implementation of BI system.

Besides mentioned acceleration of decision making BI systems bring such advantages in surveyed enterprises as an increase of decisions' efficacy at all level of management and it can be also stated on the basis of the research that BI systems have significant impact on such areas of decision making as consumer relationship management, marketing, sales, distribution, human resources management, controlling and logistics. The other selected advantages were improvement of communication among employees, increase of information which was indispensable for making a decision and improvement of reporting and different analyses realization.

V. CONCLUSION

As a conclusion it is worth mentioning that Business Intelligence systems have many advantages for the companies which implemented it, from the acceleration of decision making process at all levels of management to the facilitation of different types of analyses. As it was also mentioned BI system are a key feature for achievement of competitive advantage. All the modern technologies applied in contemporary enterprises such as BI systems, business analytics tools, data mining methods and techniques, cloud computing and big data solutions contribute to better communication within a particular enterprise, optimization and improvement of management processes.

REFERENCES

[1] T. Davenport.: Business Intelligence and organizational processes. International Journal of Business Intelligence Research, 1(1), 1-12, January-March 2010, www.igi-global.com

[2] R. Skyrius, G. Kazakeviciene, V. Bujauskas: The relationship between Management Decision Support and Business Intelligence: Developing Awareness. *Advances in Intelligent Systems and Computing* vol. 206, 2013, pp 587-598

[3] E. Turban, R. Sharda, D. Delen, "Decision Support Systems and Intelligent Systems", 8th edition, Prentice Hall, 2006

[4] Rouhani S., Asgari S., and Mirhosseini S., 2012, "Review Study: Business Intelligence Concepts and Approaches", *American Journal of Scientific Research*

[5] D. Dziembek, L. Ziora.: Business intelligence systems in the SaaS model as a tool supporting knowledge acquisition in the virtual organization. *Online Journal of Applied Knowledge Management*, Volume 2, issue 2, 2014, p.82-96

[6] D. Dziembek: Systemy Business Intelligence w modelu SaaS w działalności małych i średnich przedsiębiorstw. <http://www.ptzp.org.pl>

[7] E. Turban, R. Sharda, J.E. Aronson, D. King: "Business Intelligence. A Managerial Approach", Pearson, New Jersey 2010

[8] C. Olszak, E. Ziemia: Critical Success Factors for Implementing Business Intelligence Systems in Small and Medium Enterprises on the Example of Upper Silesia, Poland. *Interdisciplinary Journal of Information, Knowledge, and Management* Volume 7, 2012. <http://www.ijikm.org/Volume7/IJIKMv7p129-150Olszak634.pdf>

[9] Yogev N., Even A., Fink L.: How Business Intelligence Creates Value: An Empirical Investigation. *International Journal of Business Intelligence Research*, 4(3), 16-31, July-September 2013, www.igi-global.com

[10] J.C. Leontiades: *Managing the Global Enterprise. Competing in the information age.* Prentice Hall, London 2001, pp. 146-155

[11] Redwood capital sector report: Business Intelligence report, <http://www.redcapgroup.com> April 2014

[12] Enterprise Iron: large international chemical company standardizes data warehouse and business intelligence solutions. <http://www.enterpriseiron.com/large-international-chemical-company-standardizes-data-warehouse-and-business-intelligence-solutions/>

[13] G. Miller, D. Brautigam, S. Gerlach: *Business Intelligence Competency Centers. A team approach to maximizing competitive advantage.* Wiley&Sons, New Jersey 2006, pp. 15-34

[14] W. Eckerson: Business-Driven BI: <http://www.beyerresearch.com/study/16441>

[15] P. Dindigal: Healthcare Business Intelligence: Saving lives through enhanced information. Satyam Healthcare Practice. <http://www.himss.org/files/HIMSSorg/content/files/Satyam021109.pdf>

[16] A. Chluski, L. Ziora: The Possibilities of Business Intelligence Systems Application in Polish Hospitals. *Current Problems of Maintenance of Electrical Equipment and Management.* Monograph. Scientific Editors Michal Kolcun, Lech Borowik, Tomasz Lis. Technicka Univerzita v Kosciach. 2014, p.245-254

[17] L. Ziora: The role of Business Intelligence systems in the decision making process of international enterprises, Phd thesis, 2011

Leszek Ziora Ph.D., employed as Assistant Professor at Czestochowa University of Technology, the Faculty of Management, Business Informatics Department. He is the author of over 30 papers published in domestic and international journals. His scientific interests include Business Intelligence systems, data security in computer networks, big data and cloud computing solutions, application of linguistics in management. He is member of Scientific Association of Business Informatics and International Association of Engineers.

The concept of a model of the separation of the user interface layer from the database layer in B2B system

M. Łobaziewicz

Abstract - The use of the Internet and web applications in business and the implementation of processes in the B2B model causes that there is a growing demand for more and more effective methods for their creation. New design patterns and specifications appear, and standards and guidelines are being established which standardize activities carried out with the use of the Internet. Each information system is based on certain patterns. These patterns are associated with different layers of information systems and a logical division of tasks in the system. Despite the fact that patterns are independent or loosely associated with the adjacent layers of information systems, they are responsible for certain functionalities, which, as a whole, form the architecture of computer systems. B2B systems are also characterized by multilayeredness. At the level of a design of each layer of a B2B system appropriate patterns must be used. The use of such solutions is to achieve independence of the database engine from the system and the possibility of extending the functionality of the user interface in each created application.

The purpose of this article is to present the results of research and recommendations concerning the separation of the user interface layer from the database layer and identifying the main sources of data that power B2B data exchange standards in this system.

Keywords - B2B system, design pattern, layered system architecture.

I. INTRODUCTION

The complexity of business processes that are implemented by companies in cooperation with their business partners located in different parts of a country or the world and high demands placed by them mean that modern B2B systems are advanced IT solutions that make use of web applications equipped with various functionalities. Time pressure and high competition between companies is so strong that, for the purpose of rapid production of applications, ready-made architectural patterns are used, which, as it is in the products or services, are continually developed and replaced by new. Patterns are associated with different layers of information systems. Patterns often describe and even define them. Layers are logical divisions of tasks in IT systems responsible for specific functions, and they must communicate with each other, even though they are logically separated from each other and loosely connected with the adjacent layers.

Given the above, the aim of this article is to present the outcome of research and recommendations concerning the separation of the user interface layer from the database layer and identifying the main sources of data that power B2B data exchange standards in this system.

The study was carried out as a part of a project "Development of a modern and advanced B2B system based on Internet technologies as a result of research and development works" implemented by OPTeam SA financed from the Regional Operational Programme 2007-2013 of Lubelskie Voivodeship.

II. RESEARCH METHODOLOGY

Studies on the separation of the user interface layer from the database layer in order to identify the main sources of data that power the B2B system and data exchange standards in a B2B system have been preceded by the following research tasks:

- 1) Research concerning functionalities of B2B system models based on internet technologies and the development of methods of their standardization; and,
- 2) The development of new standards for B2B system functionality and the development of a model of their dependence on each other.

Research results presented in this article are the outcome of the implementation of the third phase of the project that should lead to a situation that, during the prototype creation phase, it would be possible to simultaneously work on dependent and at separated B2B system functionalities. The first purpose is to make the system engine independent from database used, and the second is to make it easy to extend the functionality of the user interface. As part of the research works, architectural standards were analysed, patterns were used in designing the users' interfaces, the B2B system was used to power models from different data sources, standards were considered for external data sent to the system and standards for data (concerning documents, payments, customer data) collected from the system in the most commonly used formats. As part of the research works on the interface layer, recommendations were developed on the choice of technological standards of design and the implementation, application layers in the application, data exchange formats, and information generated by various management systems that will work together via the B2B platform as well as the information flow.

III. LAYERED ARCHITECTURE IN DESIGNING B2B COMPUTER SYSTEMS

From the perspective of logical application architecture, each IT system is treated as a set of cooperating layers, each of which presents the specific nature of the services [1]. Layers constitute a logical division of tasks in the system. One of the most popular models of application layering is to divide it into three layers: the presentation layer (the user), the middle (business) layer, and data layer (database).

This classification allows the identification of the types of services present in each system, ensures its proper segmentation, and determines defining interfaces between the layers. During the each layer implementation, segmentation allows a selection of specific components of the architecture and the design components and the creation of the application easier to use and maintain during operation.

3.1. The Presentation Layer

The presentation layer takes the form of a graphical user interface. Its main function is to convert data from a human-readable form to a form acceptable by the system and vice versa. It is responsible for the interaction of the user with the system and constitutes a bridge that gives access to the business services layer. The traditional presentation layer is associated with interactive users' service, whereas services implemented in this layer are responsible for the pre-processing of data from other systems without the visible use of the user interface. Typically, in the presentation layer, authentication, and authorization processes occur, and their progress is dependent on their type and the user.

Today's presentation layers are so expanded that the correct interface design is a big challenge for a designer. The lack of having appropriate design patterns may, in the long term, result in a situation that any interface change will imply the interface code refactoring, which is rather complicated from the point of view of development works.

A properly designed presentation layer is characterized by the independence of the interface from the way of displaying it. In other words, the graphical user interface should be "unaware" of the data that it displays.

Another feature of the presentation layer is the independence of the graphical interface from information technology. In practice, this principle is not always fulfilled. In the theoretical approach, architecture of this layer should allow an unconstrained change of technology, but, in fact, designing so flexible layer is a very difficult and time-consuming task; therefore, designers resign from doing that. A well-designed presentation layer contains the elements of support for testing purposes. In an autonomous view, automated tests are practically impossible. It is, therefore, necessary to use Application Programming Interface (API) that allows the user to simulate a precise user action from the code level, e.g. clicking on a button. The key is to create an abstraction between view and presentation logic.

The Microsoft .NET platform has a library of Windows Forms controls and a library of ASP.NET Web Forms controls that are similar to each other. Thanks to it, the design of the user

interface (presentation layer) in web applications is similar to the visual design of the desktop web forms. In both cases, the controls are built by the use of mouse and configured in the properties window, and the application dynamics is determined by means of the event-driven study. An event-driven model, which performs well in window forms, is not natural for web applications. Therefore, an ASP.NET MVC (Model-View-Controller) tool has been created, which is a bit more complicated for programmers, as it requires a good knowledge of web-based tools (JavaScript, JQuery); however, it uses a good, for this type of application, design patterns and allows the programmer a full control of the HTML code sent from the server to the browser. Thus, while deciding to choose between ASP.NET Web Forms and ASP.NET MVC, one shall specify expectations from the application. ASP.NET MVC allows one to control more precisely all that is happening in the application. It is important for the B2B system because the key feature is its performance. As a rule, the B2B system should be able to develop in many, often unpredictable, directions, because its users represent different companies cooperating on a single platform, and they have different, often unusual needs.

3.2. Business Layer

This layer is often referred to as a business logic layer or the core of the system. It implements the utility logic and communication between the data and presentation layers. It is independent from the source and the way of data delivery, as well as from the operating systems used by the companies that cooperate via the platform; thus, it ensures the stability and flexibility of the B2B system expansion.

Business logic in an enterprise is a reflection of the processes occurring in it, and a logic layer within the programmistic meaning presents it in the form of IT model. The main tasks of the business logic are the following:

- a) Implementation of business rules: these include strategies and ways of conducting business worked out by each company;
- b) Ensuring the transactionality of processing and cooperation with the database, data validation: checking the correctness of the data entered into the system in order to ensure data consistency and to avoiding the errors of the system; and,
- c) System work monitoring, including system error handling and storing system logs.

3.3. Data layer

The data layer is connected with the server on which the application data is stored. It provides access to data through standard interfaces that can be used by business services layer components. Services in the data layer are not directly accessible from the presentation layer level.

There are several approaches to the logical division of the system into layers. They are an extension of the basic division into three layers. These include the popular J2EE patterns directory (Fig. 1).

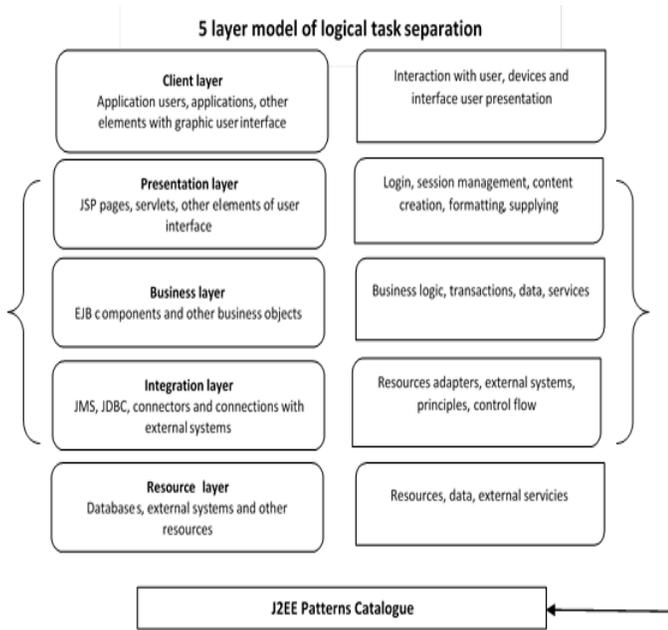


Fig. 1 Model of a logical separation of tasks in the J2EE pattern [2]

The division into layers has several advantages. One of them is the possibility of separating fragments of solutions, which allows for the independent conduct of design and development works. Another advantage of the multilayeredness is independent processing. This allows the high scalability of the solution and an optimum choice of the hardware environment for each layer. Division of the application layer into layers also provides greater security for computer systems, their transparency, and facilitates the management of the presentation layer.

IV. DESIGN PATTERNS IN A LAYERED ARCHITECTURE OF WEB APPLICATIONS

The layered approach in the design of applications is very often implemented based on “design patterns.”

Design Patterns in the information technology design have been developed on the basis of the definition by Ch. Alexander who emphasizes that each pattern describes a problem which occurs over and over again in our environment, and then describes the core of the solution to that problem, in such a way that you can use this solution a million times over, without ever doing it the same way twice [3]. Therefore, design patterns are positively verified and standardized schemes of behaviour used in the IT design of a dedicated to a given applications [4].

By using software design patterns, programs are more efficient, scalable, there is a greater possibility of their modifications, and they are less prone to errors made by programmers. Design patterns allow for more effective separation of roles of members of design and programming teamwork. There is also a significant reduction in the intensity

of network communication and an increase in access security. Knowledge of patterns accelerates solving many programmatic problems and systematises the terminology used by programmers. This saves the time required to create a well-functioning and properly build application.

The concept of design patterns was popularized by E. Gamma, R. Helm, R. Johnson, and J. Vlissides (Gang Of Four - GOF) [5], who divided design patterns into three categories:

1. Creational patterns that show how to make object more flexible: They delegate the process of creation to other classes (which is important due to the reduction of dependencies in the code), and specifying the control over the creation of objects.
2. Structural patterns show how to connect classes with each other, defining the management of the structure of objects and structures composed of objects.
3. Behavioural patterns show how to make the behaviour of the software more flexible, specify the behaviour of objects, and communication between them.

Within categories, specific patterns have been defined, each of which comprises of fixed elements: name, a description of the problem that the pattern shows, and opportunities and results of the pattern used [4].

Below, only those patterns that are used in the design of Web applications, which include B2B systems, have been discussed.

A. Model View Controller (MVC)

The MVC pattern [5] is used to create applications that have a graphical user interface. Its role is to enforce the division of the application into 3 independent layers. It forces the system architecture to be orderly. These are as follows:

- **Model:** It contains a description of the data structures and relationships between them. It contains data and provides operations not related to the user interface service, which are “business logic functions.” It constitutes the core for the functionality of the application, registers dependent views, and controls and informs dependent components of the data change. This model is an access gateway to the business layer. It stores (“keeps”) information about the current state of the interface.

- **View:** It is a screen presentation of the model. It initializes the display of information to the user, and it gets the data from the Model and implements the update. It is used to present the data and to log out of the applications, e.g., in HTML; however, the results may be presented in many formats, e.g. XML files, PDF files.

- **Controller:** This is the logic of the operation, namely, the relationships between events occurring in the system. The task of the controller is receiving, processing, and analysing the user input data. In a typical application, input data sources are the keyboard and mouse. By virtue of their function, both the Controller and the View have equal access to the data model. Thus, the controller does not allow changing the presentation layer without modifying the logic. For this reason, the controller should not contain the logic of the application, only a reference to it. In the majority of solutions, the View and Controller are linked into a single object (Document View).

In case of a web application, the Controller's role is different. All tasks are sent to the Controller, which maps them in

methods calling of the model. Then, the results of these actions are passed to the View. The tasks of the Controller also include the provision of the security of application and data validation. In a web application, it is a combination of Front Controller that supports all the tasks sent to the site by directing them to one object and an Application Controller that controls the entire application.

In ASP.NET technology, which was used for the design of the B2B system, the Controller performs tasks in response to View's tasks. The implementation of a controller in ASP.NET resembles a Page Controller design pattern, in which, for each site, there is a module that fulfils functions of its Controller. Front Controller patterns are also possible to implement. Model class is not required. The functionality of the Model can be implemented in a separate class or belong to the Controller. ADO.NET library supports the service of Models. It is not required to maintain an open connection to the database. Applications maintain a connection only while data reading or recording. Data downloaded from the database using ADO.NET is stored in the DataSet that acts as a buffer.

B. Mobile Front Controller (MFC)

The MFC pattern [6] is a model of the application layer and is often used together with the MVC pattern. It provides a common service of requests for the presentation layer components. Centralization of views management (View layer elements) of a web application user in a single object accepts users' requests. The use of a MFC pattern is recommended for web applications with complicated navigation of dynamically generated sites and for web applications that require authentication, security policies, transformations, etc.

C. Model View Presenter (MVP)

A MVP model [4] is a design pattern derived from the MVC pattern. The difference is that the Controller with MVC is a Presenter (Supervising Controller). This means that all the results of the business logic of the application are sent from the Presenter, and not, as in the classic MVC pattern, from the Model.

While comparing the MVP of MVC, it may be observed that, in MVC, the Controller is based on a request and combines adequate Models with corresponding Views. The View retrieves relevant data from the Model according to its discretion and creates an answer. While, in the MVP, the Presenter (Controller) extracts the data needed for site rendering from Models and forward them to the View.

V. SERVICE ORIENTED ARCHITECTURE

According to Garthner, Service Oriented Architecture (SOA) is an architecture of a software that is based on interface definition [8]. It creates application topologies based on services, and their implementation and calls.

IT corporations, such as IBM, Oracle, Microsoft, defines SOA as the concept of building IT systems of service-oriented enterprises as shared between two applications proven

components [9]. SOA is also defined as an architecture for business applications created as a set of independent components, organized in such a way as to deliver services, operating according to certain criteria, and supporting business processes implementation [8]. This definition shows that the SOA is neither a technology nor product, but an effective approach used in designing and integration of services.

The concept of service-oriented architecture is based on the assumption that the business logic is not a monolithic program, but it consists of many service components coordinated by a central managing application.

SOA categorizes the relationships between service providers and their recipients represented by software components implementing established complex business processes. It provides reuse of software components, encapsulation of functionality, precise definition of interfaces, and flexibility of applications created by composition method. SOA components are loosely related to each other but cooperate with each other to implement a business process.

In SOA, as the pattern of designing business-oriented heterogeneous dispersed systems, such as B2B, two types of solutions are differentiated:

1. Web Services, which are defined by a specialized Web Services Description Language (WSDL). Services are published in the Service Registry (SR). They can be remotely called via a defined interface. Available services calling may be coordinated through the mechanism of service composition, which manages the execution of business processes based on services from various systems, and the whole is made available as a new service.

Construction of new, complex network services is made through the following:

- Orchestration, defined as "an executable business process describing a flow from the perspective and under control of a single endpoint (commonly: Workflow)" [11]; and,
- Choreography, which allows the display of the business process part describing interactions with a given service outside the organization (Web services choreography mechanism is defined by WS-CDL (Web Service Choreography Description Language)) [12].

2. Representational State Transfer (REST) is a technology that enables the creation of the protocol and makes data transmission possible by means of HTTP protocol features. It does not apply an additional layer to message sending, such as, for example, SOAP. This service should be described using WSDL.

While designing applications based on SOA, it is required to understand four basic design patterns for Web Services [13]:

- Facade provides a common interface for multiple business components. It is located under the business logic layer. It is used to reduce the number of connections and dependencies between systems. The Facade is a measure of access to a complex system that presents outside a simplified and structured programming interface.
- Adapter is a service that simplifies and facilitates access to the site that provides the service. It serves the purpose of adjusting the object-oriented interfaces so that the cooperation of objects with incompatible connections would be possible. In

SOA, it solves the problems of incompatible interface prevalence.

- Proxy acts as the representative of another object in order to obtain a supervised access to the object that it represents. It functions as an interface to another object, e.g., an Internet connection, a large object in a memory, file, or another resource. Proxy manages Internet connections and provides the user with the functionality that it obtains as a client of a remote server, as if it makes it available itself.
- Controller is the mediator between the user interface and the data layer. It provides business logic. In SOA, it serves the purpose of using existing MVC applications and the encapsulation of the complex business logic of the system. As part of SOA, communication patterns are also used because services can be called by other services or programs. Rules of communication must be established for that purpose. Information about services is provided through dedicated descriptions (service descriptions) that contain the name of the service and data standards expected and returned by this service. Communication patterns include the following:

- Document Message is a pattern of a message that makes it possible for a sender to send data structure to the recipient. It is an independent unit of communication, a single object, or data structure subject to decomposition into elementary components. Document Message may be any type of message in the system.
- Request-Response is a basic synchronous communication template. G. Hohpe and B. Woolf define it an answer to the question: When an application sends a message, how can it receive a reply from the recipient [14]? In this model, the client sends a request to the service and waits for its reply. Complete interaction requires communication infrastructure.

VI. THE STRUCTURE OF DATA LAYER

The Data Access Layer is responsible for the organization of the persistent data stored on file servers or in databases, making them available to higher layers so that they could use these data without the knowledge of the structure of their storing. The heart of this layer is usually a relational database, because the implementation of which is responsible for one of many systems implementations of relative database management based on a well-known SQL standard.

Repository is used in creating data access, which is an additional layer that separates the object-oriented data access layer. It behaves like a collection in memory of isolating the character entities from business layer of the data infrastructure. It operates at the level of one class model. Usually, one single repository for a business object (Aggregate Root) is being defined. A very important feature of the repository is its independence from database. The user should use the correctly implemented repository as the ordinary class or collection of data.

Similar to the Repository, and used in service-oriented applications but at a lower level, is the Data Access Object. It stems from the observation that the data of business components must be stored in non-volatile memory (mass), in the ordinary file system, in the repository of XML files, or in databases. Each of these memory types has its own unique

API. The differences in the API cause that, in the case of a change of the system supply, it is necessary to modify the data layer components, and sometimes the business layer components. On the other hand, making the application independent from the data storage, the system requires the creation of a universal interface to access data in mass storage.

At the level of the data access layer, a Unit of Work pattern is used. It makes the access to all Repositories from one place possible. In addition, it is responsible for the management of transactions (keeps in a memory all the updates, and then sends to them to database as a single transaction), manages a list of objects that take part in a business transaction, records changes made on business objects (addition, deletion, modification), and coordinates their preservation in the database as a single transaction or withdrawal from the database in case of failure, ensuring data integrity.

VII. APPLICATION LAYERS VS PROCESS OF DESIGNING

Application layering has a significant impact on the process of design and execution of development works. There are rules to be followed during both of these processes. Object Oriented Programming SOLID (OOP) defines five principles of object-oriented programming, which allow, among other things, to react quickly to the change of requirements in the project. These include the following [14]:

1. Single Responsibility Principle (S): According to the principle, "there should never be more than one reason for a class to change." Class or module should exist for one purpose only, and they should have to perform a single activity.
2. Open / Closed Principle (O): In accordance to this principle, software entities (classes, modules, functions, etc.) should be open for extension, but closed for modification. Openness for extension means, in the case of a B2B system, a design of classes that new functions are possible to be added. Closed for modification means that after class designing, one should never modify it, except to correct errors. Despite the fact that these two principles may seem contradictory, building correct classes, modules, or functions allows adding new functions without editing and modifying the existing source code. One should be able to create new classes that do the same thing differently without changing the base functionality.
3. Liskov Substitution Principle (L): It is based on the assumption that "Functions that use pointers or references to base classes must be able to use objects of derived classes without knowing it." While making the foundations for a given software, the designers build hierarchies of classes, based on which, the extensions for these classes are formed. This principle boils down to the fact that, in every place in which base class is used, it is also possible to use derived classes, and this fact should not affect the functionality implemented by the application.
4. Interface Segregation Principle (I): This rule specifies that the user should not be forced to implement interfaces that he/she does not use. According to this principle, one should divide expanded interfaces into smaller ones. Clients should not be forced to depend upon interfaces that they do not use.

5. Dependency Principle Invention (D): Under the principle of "high level modules should not depend upon low level modules. Both groups of modules should depend on abstractions." This rule specifies high-level modules independencies from specific mechanisms that belong to low-level layers. By applying SOLID principles, one is able to create code in which it is easy to make changes and to respond to changes of requirements.

VIII. COOPERATION WITH WEB SERVICE

Separating the data sources from the client interface in a form of a Web Service layer causes that the B2B system can be powered by any external data with a fixed format. These may be complete data that ensure all system functionalities or partial data that power only a certain part of it. By using Web Service, which, in technical terms, is a technology of a distributed processing based on web technologies, it provides high universalism of use in case of the dedicated B2B system, while also allowing for integration with any external systems. These may be both web applications and desktop programs, but also more and more popular and widespread mobile programs.

Figure 2 shows an exemplary model of the system structure based on Web Service and a customer access layer through a dedicated interface (powered by data from Web Service).

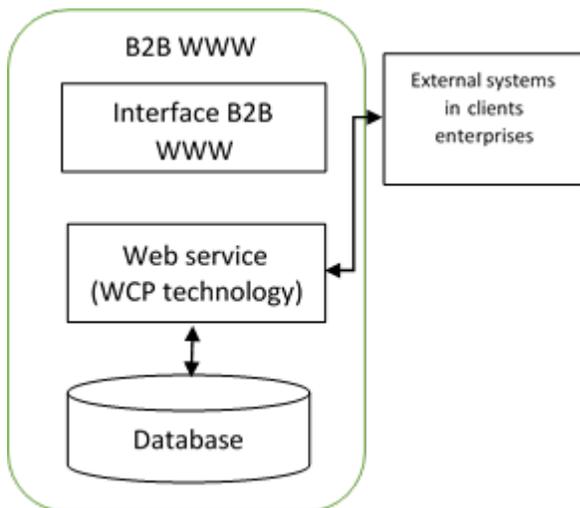


Fig. 2. Model of the separation of B2B system using Web Service

By applying a layered structure and Web Service, the access to external data is separated the from the data layer of the B2B system. Access to these data should be adequately protected so that confidential data is transmitted only to authorized users (e.g. data concerning payments, settlement prices of goods that include discounts for particular customers). In practice, such a solution enables the possibility to make the company's offer available to other companies, so that they do not have to store the information in their databases. This has several important advantages, including the lack of mechanisms of replication, constant access to the latest data, and the lack of expenditures

to increase disk space (less data locally stored). The mechanism not only provides uploading of current data from B2B system, but it also, if procedures are properly implemented and made available, a record of orders, complaints, or service requests. It is also possible to divert the situation and power the access layer (B2B Web Interface) by data collected from one or more external Web Service's of other companies. For example, a popular method is to incorporate in B2B data information on packages or shipment tracking. Shipping companies make web services available that allow a person to get data about a shipment and place such data in other web solutions. This results in a system that is more uniform and clear (no redirecting to external sites).

The use of Web Service technology in B2B solutions seems to be very beneficial and provides more flexibility in the later development of the platform.

IX. ARRANGEMENT OF LAYERS AND HARDWARE STRUCTURE

The division of application into logical services layers does not necessarily have to mean exactly the same division of the physical model of the application. The completely layered application may be placed on a single server or distributed among the physical devices if needed.

Using a few servers may minimize the cost of the system. In a situation that it is not possible to have access to a few devices, the designers of the application shall place the data access services and business services on the same server (Figure 3). The application communicates with the database server (SQL Server) through one of the layers (data layer).

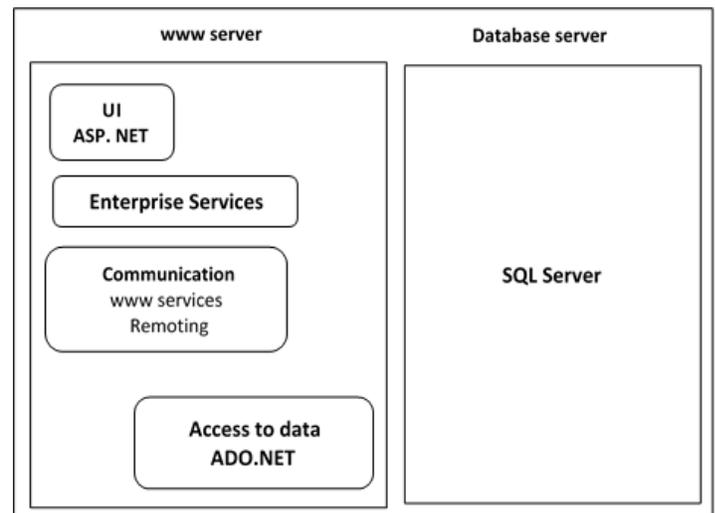


Fig. 3. The web server as an application server

Distribution of particular application layers may be arbitrary - from a simple form, one server to distribute each layer on a separate server. Below is a diagram of separating a "remote application layer" that enables one to isolate web layer from application layer.

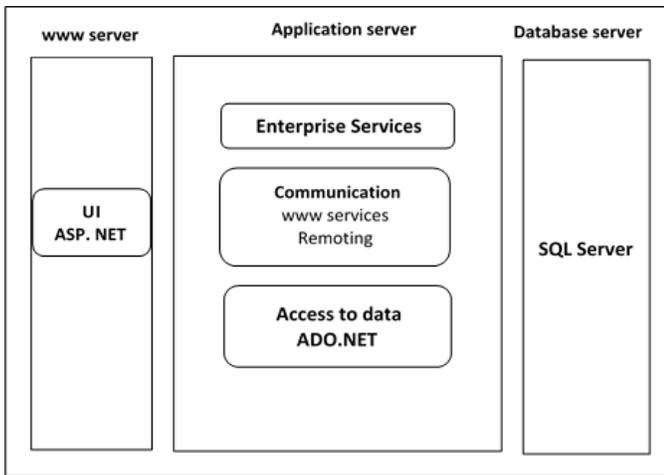


Fig. 4. Introducing a remote application layer that isolates the web layer from the application layer

The advantage of the layered design of the application is, among other things, the possibility of an arbitrary arrangement of layers in client's hardware structure. The division means the possibility of a distribution of application components that work with the optimum efficiency. It also provides the highest security standards.

X. B2B SYSTEM POWERING BY EXTERNAL DATA. PRACTICAL POSSIBILITIES OF USING

One of the most useful functions of B2B systems is the possibility to automate certain tasks. For example, it is possible to build automated orders entry functionality to the system by importing external files in any specified format. Companies that make use of B2B systems have their own systems, and they are often able to export their orders created in the ERP system to external files. Therefore, the desired function would be the possibility of their introduction into the B2B system in which orders are created by only indicating such a file.

Remembering that B2B systems automatically charge the appropriate discounts for a logged in customer, the amount of data needed to automate the orders entry is small and it is limited to the following basic fields:

- Product code - a unique code of product, recognizable in the B2B system, and
- Quantity - quantity of the product being ordered.

Considering the mode of operation, it is possible to distinguish two possible scenarios to automate orders entry:

- As a ready-made order in the ERP system, and
- As goods kept in a shopping cart.

Studies have shown that, from the functional side, the second method provides greater possibilities and is more universal. Apart from products automatically imported from the file, one is able to manually add additional items to the shopping cart and place an order. For the purpose of the creation of a prototype of B2B system, implementing imported goods as next items in shopping cart of an order was chosen.

Furthermore, next order processing undergoes standard procedures of ordering, that is going through a stage of the order confirmation and final placing of an order.

Another very good example of B2B integration with external data is the use of solutions common for shipping companies, which usually create the possibility of viewing data on shipment tracking. This type of solution, integrated with B2B system, means that the client has full information about the status of the shipment with its details that shipping companies provide in one place. Companies provide this type of information as a Web Service. They enable the integration of data also in the opposite direction, i.e. through recording the data to the systems of shipping companies. These solutions are built in accordance with applicable standards. On the websites of shipping companies, there may be found information on how to integrate external systems with these services.

XI. PROJECT RECOMMENDATIONS

The designed B2B system should meet the market standards of design and construction of this type of application. Because of the adopted by OPTeam S.A. technology of programming on the ASP.NET platform, it is suggested that the latest, in this regard, Microsoft solutions are most suitable. They provide support for the described design patterns, and cooperate with accepted standards of data transmission. The project should meet the assumptions of work dedicated to the specifics of the modern B2B system and support dispersed work. It is recommended to separate three main layers:

- Presentation layer,
- Business layer,
- Data layer.

The use of Web Service in WCF technology, which is a combination of Web Service XML and .NET Remoting features, provides the following:

- High scalability of the created system, and
- The possibility to download data by external clients through various sources (mobile, web, desktop programs), without limiting to the use of the system by the company that owns the platform.

Separating logical layers and creating Web Service opens the system for work with an unlimited number of programs, which, in turn, enables the integration between companies. Web Service allows one to publish the services outside and the exchange of information, which often limits the use of sometimes "defective" human factor. It also provides appropriate architectural and efficiency solutions that are important if an increasing number of users is being observed.

Based on OPTeam SA experience, customers who report the demand for B2B systems have the potential to assign at least 2 separate dedicated servers for the solution; hence, the separation of adequate layers between two or more hardware devices would cause an increase in its capacity and gives an opportunity to strengthen these layers, which will require it. This is particularly important at the level of data access layer, which should provide the necessary speed of response to inquiries from the logic layer. One should keep in mind that

companies use many other solutions that overload the database and the B2B system indirectly. Diagnosing the "bottlenecks" of a computer system and possibilities to support them by more efficient hardware speaks in favour of a multi-layer solution. Comparison or benchmark tests are useful in this respect, and they should be carried out as part of the next stage of the project of creating a prototype of B2B system. Research carried out as part of the project on Web Service technology suggests two possibilities: the use of Java (JEE / J2EE) and Microsoft.Net.

research interests are focused on management information systems, e-business systems, business and information processes, High-Tech innovations in enterprise management. For years, she has been cooperating with business, designing advanced solutions for the improvement of management systems. In OPTeam S.A. she is the scientific Project Manager in the R&D project "Development of state of the art and advanced B2B system based on Internet technologies as a result of research and development works". The project is co-financed by EU funds in frame of Regional Operational Program for 2007-2013 of Lubelskie Voivodship.

REFERENCES

- [1] J.D. Meier, A. Mackman, M. Dunner, S. Vasireddy, *Building Secure Microsoft ASP.NET Applications*, 2002.
- [2] D. Alur, D., Malks, J. Crupi, *Core J2EE Patterns: Best Practices and Design Strategies*, 2nd ed., Prentice Hall & Sun Microsystems Press, 2003.
- [3] Ch. Alexander, S. Ishikawa, M. Silverstein, M. Jacobson, I. Fiksdahl-King, S. Angel, *A Pattern Language*, Oxford University Press, NewYork, 1977.
- [4] E. Gamma, R. Helm, R. Johnson, J.Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison Wesley, 1994.
- [5] M. Fowler, D. Rice, M. Foemmel, E. Hieatt, R. Mee, R. Staffor, *Patterns of Enterprise Application Architecture*, Addison Wesley 2002.
- [6] K. Żydzik, T. Rak, *C# 6.0 i MVC 5. Tworzenie nowoczesnych portali internetowych*, Helion, 2015, p. 127.
- [7] T. Erl, C. Gee, J. Kress, B. Maier, H. Normann, P. Raj, L. Shuster, B. Trops, C. Utschig-Utschig, P. Wik, T. Winterberg, *Next Generation SOA. A Concise Introduction to Service Technology & Service-Oriented*, Prentice Hall & Pearson PTR, 2012.
- [8] J. Hoon Lee, H. J. Shim, K. K. Kim, Critical Success Factors in SOA Implementation. An Exploratory Study, *Information Systems Management*, 27:123–145, 2010, pp.124-125.
- [9] J. Łagowski, *SOA – ideology, not technology*, XV Conference PLOUG, Kościelisko, 2009.
- [10] H. Haas, A. Brown, Web Services Glossary: Available: <http://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/>.
- [11] W3C Web Services Choreography Working Group: Available: <http://www.w3.org/TR/ws-cdl-10/>.
- [12] J. Fronckowiak, *SOAB est Practices and Design Patterns Keys to Successful Service - Oriented Architecture Implementation*, White Paper Published by Oracle Corp, 2009.
- [13] G. Hoppe, B. Woolf, *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA ©2003.
- [14] R. C. Martin, M. Martin, *Agile Principles, Patterns, and Practices in C#*, Prentice Hall & Pearson Education, 2006.

M. Łobaziewicz, PhD in Economics, Faculty of Social Sciences, Institute of Economics & Management, John Paul II Catholic University of Lublin. Her

Implementing the green transport strategy using Balanced Scorecard and Analytic Network Process

D. Staš, R. Lenort, P. Wicher, and D. Holman

Abstract—The paper presents a conceptual framework for supporting a green transport strategy implementation in industrial companies and supply chains. Balanced Scorecard and Analytic Network Process were used as the methodological basis. The framework contains five main steps – selection of green transport strategy, selection of Balanced Scorecard approach, specification of green transport strategy, prioritization of Balanced Scorecard measures, and evaluation of green transport strategy reaching. For each phase are defined fundamental principles and recommended tools. The verification of the designed conceptual framework was performed on a real supply chain of the European automotive industry.

Keywords—Analytic Network Process, automotive industry, Balanced Scorecard, green transport strategy.

I. INTRODUCTION

IN recent decades, the performance of economic and non-economic activities has required them to be friendly with the environment [1], [2]. This proactive approach to addressing and eliminating the negative environmental impacts from transport processes is called Green Transport (GT).

Transport is one of the areas having considerable potential within the scope of the green strategy implementation, since it has significant negative impacts on the environment [3]. They include, primarily, the emissions of CO and CO₂ and other exhaust gases, noise, and, last but not least, congested transport infrastructure.

The current goals of GT are now focused on reducing the fuel consumption (which is closely linked to cutting CO₂ and other exhaust gases), reducing noise, reducing the transport costs, reducing traffic jams and, ultimately, on complying with the legislative restrictions. An active and effective solution of the issues of GT must be seen not only as a challenge, but especially as an opportunity offering the possibility of significant competitive advantage, improving the image of the company in the eyes of the customers, region, state and the general public.

This work was supported by the project of the SKODA AUTO University Internal Grant Agency No. SIGA/2014/01.

All authors are with Department of Logistics and Quality Management, ŠKODA AUTO University, Na Karmeli 1457, Mladá Boleslav, Czech Republic (e-mail: stas@is.savs.cz, lenort@is.savs.cz, wicher@is.savs.cz, holman@is.savs.cz).

The aim of the paper is to design conceptual framework for supporting a Green Transport Strategy (GTS) implementation in industrial companies and supply chains using Balanced Scorecard and Analytic Network Process.

II. METHODOLOGICAL BASIS

A. *Balanced Scorecard*

Full name of the tool is System Balanced Scorecard enterprise (BSC). It is a method of management that creates a link between strategy and operational activities with an emphasis on performance measurement developed by Kaplan and Norton [4].

By using the BSC, the strategy and vision of the company can be converted into performance measures that include both outcome measures and the drivers of these measures [5]. For a strategy to be successful, it needs to consider financial ambitions, processes to be improved, markets served and the people in the organization that implement the strategy [6]. The BSC uses all these perspectives by considering both internal and external aspects [7]. Every perspective should contain four different sections: objectives, measures, targets and initiatives. For employees to be able to act upon the organization's vision, translating the strategy and mission of the company into objectives is the first step in the creation of each perspective.

Strategies like “an empowered organization” is hard to implement in practice and senior executives should therefore create understandable and actionable objectives, along with defined measures to keep track of the progress of reaching each goal [8]. Each measure should then be associated with a target (a short-term goal) that works as a milestone to assist in evaluating the progress of each objective. The last column in each perspective should be initiatives, describing actions that should be undertaken by the firm to reach each objective.

B. *Analytic Network Process*

The Analytic Network Process (ANP) is multistage decomposition method used to solve decision-making problems involving more than one criterion of optimality developed by Saaty [9]. The basic idea is to create a decision-making network and the subsequent evaluation of importance of the individual links among the interconnected elements. These evaluations are represented by weights, which are

determined on the basis of pair comparison or by normalizing direct measurements. The ANP does not limit human understanding and experience to force decision-making into a highly technical model that is unnatural and contrived. It is in essence a formalization of how people usually think, and it helps the decision-maker keep track of the process as the complexity of the problem and the diversity of its factors increase [10].

III. CONCEPTUAL FRAMEWORK DESIGN

Designed conceptual framework includes the following steps: Selection of green transport strategy, Selection of Balanced Scorecard approach, Specification of green transport strategy, Prioritization of Balanced Scorecard measures, and Evaluation of green transport strategy reaching.

A. Selection of Green Transport Strategy

The key step of the conceptual framework is selection of an appropriate green transport strategy. For that purpose, authors of the paper offer GTS matrix, which is shown in Fig. 1.

Green effect	high	I. Ecological	I. Ideal
	low	I. Ineffective	I. Economic
		high	low
		Costs	

Fig. 1 Green transport strategy matrix

The GTS matrix is based on the following criteria: (1) Expected green effect after the GTS implementation – low or high; (2) Estimated cost of the GTS implementation – low or high; (3) Responsibility to decide on the GTS implementation in the given company: I. In the responsibility of the implementers or II. Limited responsibility of the implementers (e.g. within the responsibility of another company department or corporation).

The result are four main GTSs: (1) Ideal – high green effect can be achieved at low costs or even cost savings; (2) Economic – only a limited green effect can be achieved at low costs or even cost savings; (3) Ecological – incurring high costs will achieve a high green effect; (4) Ineffective – incurring high costs brings only a limited green effect.

Ideal GTS is generally used in companies and supply chains, which start with green politics. Economic and

Ecological GTSs are implemented when Ideal GTS is depleted. Ineffective GTS should not be used at all. At the same time, green initiatives within the direct responsibility of the implementers are preferred in frame of the selected main GTS.

B. Selection of Balanced Scorecard Approach

According to Butler et al. options for incorporating sustainability/green into the BSC include: (1) Adding a fifth perspective to the BSC; (2) Developing a separate sustainable/green BSC; (3) Integrating the measures throughout the four perspectives [11].

Adding a fifth perspective to the BSC may be the simplest approach. For example, Kurien and Qureshi propose Environment perspective with three indexes: environment, social, and economic [12]. It could provide more visibility but not necessarily increased importance to the sustainability/green ability aspects of corporate management. Isolating sustainability/green measures in a separate perspective might weaken environmental initiatives by not providing clear ties to the other perspectives and to corporate strategies.

The strength of developing a separate sustainable/green BSC is the fact that a sustainable/green BSC can be used to implement a sustainability/green strategy [13]. However, the free-standing nature may fail to help the company tie sustainability directly into corporate strategy. There are two possibilities: (a) to use four original perspectives with completely new sustainability/green measures (see e.g. [14]) or (b) to develop new sustainable/green perspectives. For instance, Hsu et al. propose perspectives as follows: Sustainability, Stakeholders, Internal business processes, and Learning and growth [15].

Integrating new measures throughout the existing four perspectives has the advantage of allowing the measures to be seen as fundamental to day-to-day operations. Integration indicates that management recognizes there are cause and effect linkages between corporate strategies and sustainability/green efforts.

Authors of the paper suggest to use the second approach with four original perspectives for GTS implementation on the company/supply chain transport level and the third approach for company/supply chain level. Designed Green Transport Balanced Scorecard (GTBSC) is shown in Fig. 2.

There are two basic differences in comparison with traditional BSC: (1) Only green measures are taken into consideration; (2) In addition to target values there are threshold and real values. Thresholds represent minimum accepted values of the measures. Realities describe real values of the measures.

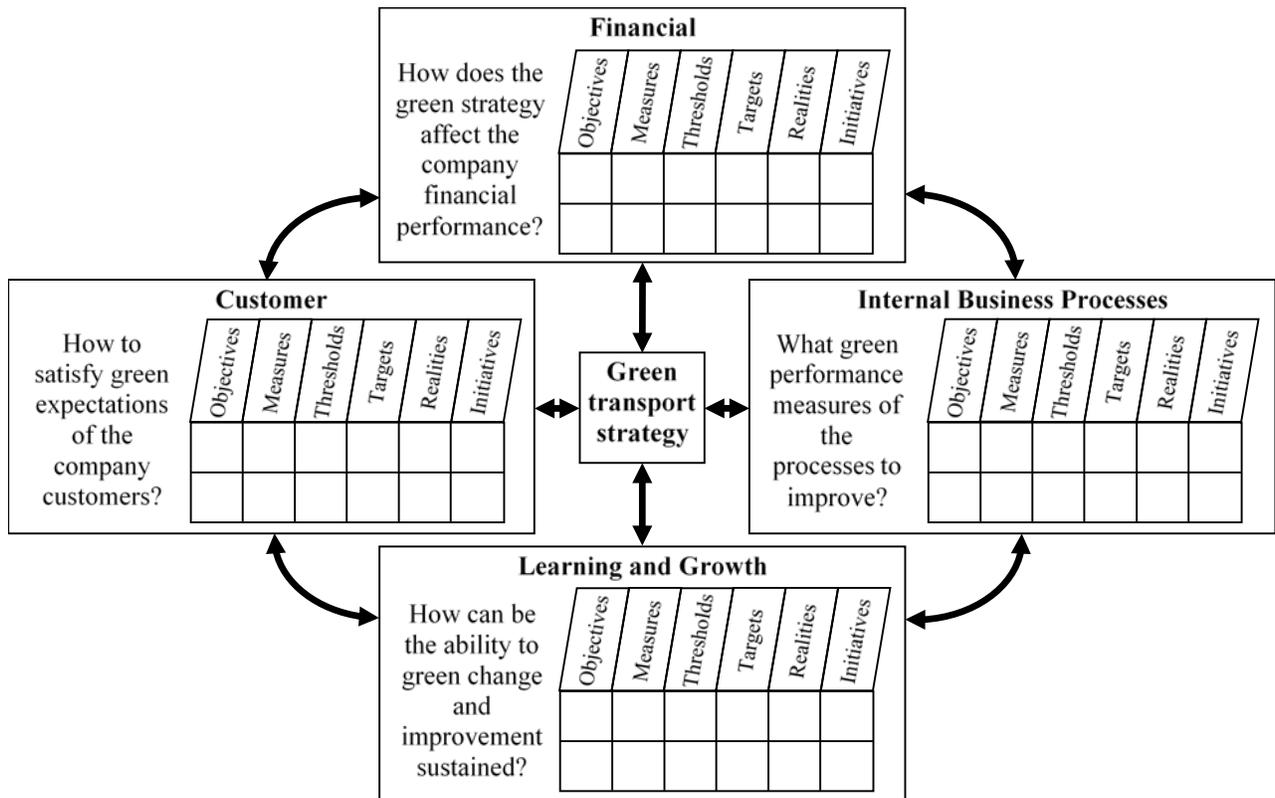


Fig. 2 Green Transport Balanced Scorecard

C. Specification of Green Transport Strategy

Specific green objectives, measures, thresholds, targets, and initiatives are determined in this step according to the selected GTS and the contemporary green transport level in the given company/supply chain. Authors of the paper designed conceptual framework for assessing the green transport level in industrial companies and supply chains for that purpose. The framework contain 30 general green best practices (initiatives), which are divided into four areas: (1) Strategy – practices creating the basis of a successful application of other best practices or they have the character of supply chain structural changes; (2) Management – practices focused on planning and subsequent execution of transport; (3) Technology – technical innovations of the means of transport, equipment, ICT systems and packages; (4) Staff – practices whose motive power is represented by the people and their skills. [16].

D. Prioritization of Balanced Scorecard Measures

The task of this step is the creation of a system for measurement of the reaching the selected GTS. It is based on the assignment of weights to the four perspectives and their measures. Authors of the paper suggest the ANP method for that purpose, because there are significant dependences between the perspectives and also their measures.

Measures with the highest weight should be incorporated to the existing company/supply chain BSC to ensure the unity between a company/supply chain strategy and the GTS.

E. Evaluation of Green Transport Strategy Reaching

A real values of the selected measures are collected during this step. Using ANP method, level of the selected GTS reaching can be calculated. The evaluation of the results may include: (1) Comparison of the calculated value with the overall threshold and target values; (2) Inclusion of the calculated value into the pre-defined categories (unacceptable, bad, good, very good, excellent GTS reaching); (3) Analysis of the trend if the evaluation of the GTS reaching is performed repeatedly. If there is the unsatisfactory GTS reaching, it is desirable to focus on the perspectives and measures with the highest weight.

IV. CASE STUDY

The verification of the designed conceptual framework is performed on a real supply chain of the European automotive industry. The GTS implementing took place in a company which is incorporated in a multinational corporation. Given the sensitivity of the used data, this section presents only an illustrative case study.

A. Selection of Green Transport Strategy

GTS is related to inbound, internal, and outbound transport, which is planned and controlled by the company. The Ideal GTS is preferred in this case study.

B. Selection of Balanced Scorecard Approach

As the GTS is only a partial strategy of the company

(Company strategy → Green strategy → Green logistics strategy → GTS), the designed GTBSC appears as an appropriate tool for GTS implementation.

C. Specification of Green Transport Strategy

Specific green objectives, measures, threshold, target and real values, and initiatives for each perspective sums up Table 1. Evaluation of reaching the objectives and their measures is carried out on the annual basis. Threshold, target and real values of the measures F1, P1, L1, L2, and C2 are expressed as annual change in per cents. Only objectives and initiatives related to the Ideal GTS in the responsibility of the company were selected.

D. Prioritization of Balanced Scorecard Measures

A network structure, which expresses dependences among the perspectives and measures is shown in Fig. 3. Orientation of the arrows determines the type of the dependences. SuperDecision software was used for the application of the ANP method (see Fig. 4). The software was written by the ANP Team, working for the Creative Decisions Foundation. There are subnets at each measure, which are used for assignment of the threshold, target and real values.

Global weights of the measures obtained using the SuperDecision software are shown in Fig. 5 in the “Limiting” column. The most significant measures are F2: High return on investments in green projects, L1: Increasing the green knowledge, and L2: Increasing the green innovativeness of logistics staff. Significance of the F2 measure corresponds to the Ideal GTS, which prefers such green initiatives that are related to low costs or with cost savings. Significance of the L1 and L2 measures is given by logic and dependences of BSC method. Perspective Learning and Growth affect positively all other perspectives. The successful implementation of GTS in long term horizon depends on high skilled and innovative logistics staff. These three measures should be incorporated into the company BSC to ensure the unity between a company strategy and the GTS.

Lower global weights were obtained in case of the C2, F1, and P1 measures, The C1 measure was evaluated as insignificant.

Table 1 Specification of the selected GTS

Perspectives	Objectives	Measures	Units	Thresholds	Targets	Realities	Initiatives
Financial	Transport cost saving	F1: Transport costs / Produced cars	EUR/car	1%	3%	2.5%	4 - Efficient system of green transport monitoring indicators
	High return on investments in green projects	F2: (Net project benefits / Project costs) * 100	%	0	20	5	
Internal Business Processes	Decreasing the CO ₂ emissions	P1: CO ₂ emissions / Driving distance	g/km	1%	3%	4%	6 - Logistics service providers with implemented green politics
Learning and Growth	Increasing the green knowledge	L1: Green training hours / Number of logistics staff	hours	25%	50%	30%	14 - High transport capacity utilisation 27 - Eco-efficient motivation system for company logistics staff
	Increasing the green innovativeness of logistics staff	L2: Number of successful green innovations / Number of logistics staff	pcs	0%	30%	25%	
Customer	Increasing the green image of transport	C1: Number of positive evaluation in a survey	%	60	90	75	29 - Green training of company logistics staff
	Reducing the local environmental impacts	C2: Number of arriving trucks / Produced cars	pcs	1%	4%	3%	

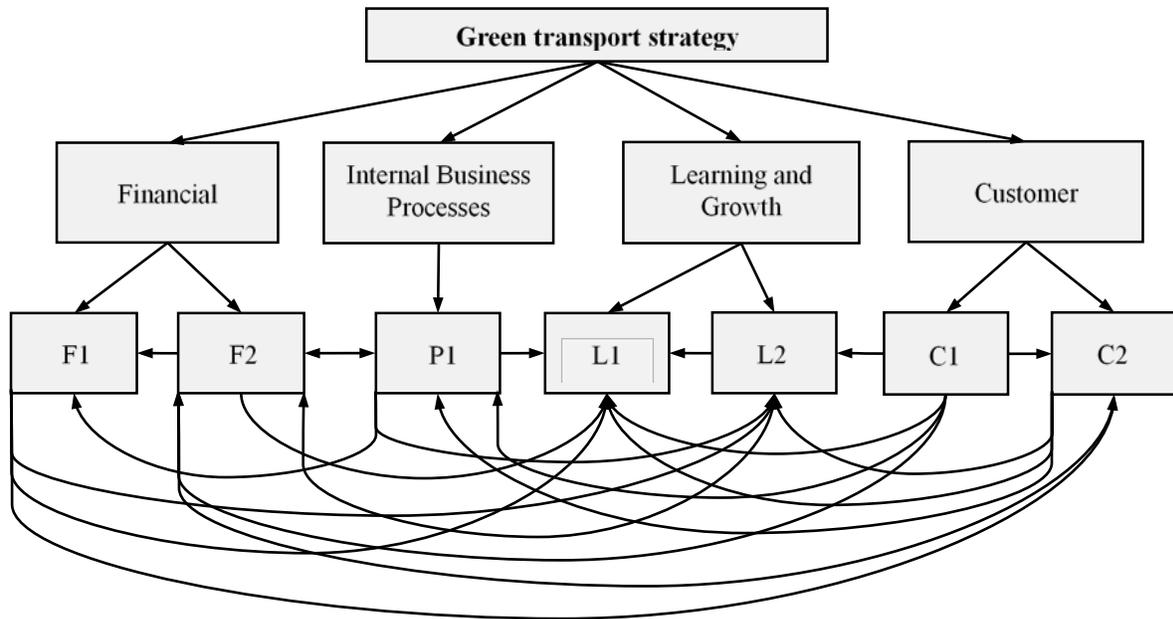


Fig. 3 Network structure of the designed GTBSC

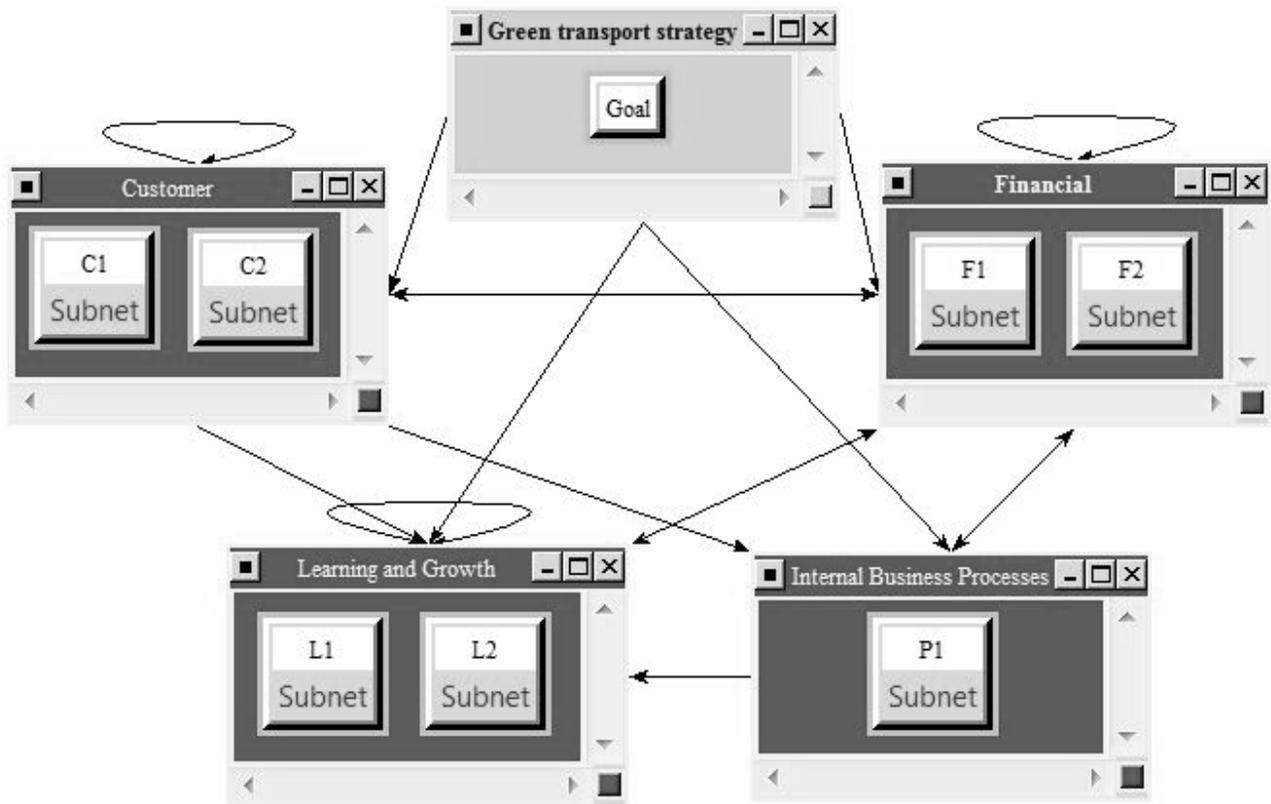


Fig. 4 Network structure in SuperDecisions software

Name	Normalized by Cluster	Limiting
C1	0.00000	0.000000
C2	1.00000	0.106348
F1	0.42796	0.154705
F2	0.57204	0.206793
Goal	0.00000	0.000000
P1	1.00000	0.106953
L1	0.51370	0.218425
L2	0.48630	0.206776

Fig. 5 Global weights of the measures

E. Evaluation of Green Transport Strategy Reaching

The main result of the evaluation step using the SuperDecision software is shown in Fig. 6. The present state of reaching the green objectives and GTS is 69%.

Name	Graphic	Ideals
Realities		0.689128
Targets		1.000000
Thresholds		0.219990

Fig. 6 Evaluation of the GTS reaching

Thanks to the overall threshold value, inclusion of the calculated real value into the pre-defined categories can be done. A value less than the overall threshold value means the GTS is implemented in an unacceptable way. On the contrary, a value greater than the overall target value means excellent GTS reaching. Interval between the overall target and threshold values can be divided into three categories (see Table 2). A suitable correction initiatives must be prepared for each category.

Table 2 System for evaluation of the GTS reaching

Category	Interval	Corrections
Excellent	> 1.00	Unneeded
Very good	0.75 – 1.00	Small
Good	0.49 – 0.74	Large
Bad	0.22 – 0.48	Principal change of GTS
Unacceptable	< 0.22	Total change of GTS

Present state of the evaluated GTS implementation is good that is why large correction initiatives must be planned and realized.

V. CONCLUSION

The presented case study has demonstrated the viability of the designed conceptual framework for supporting a GTS implementation in industrial companies and supply chains. Future works will be oriented on adding the fuzzy approach to the framework.

REFERENCES

- [1] A. Saniuk, D. Caganova, M. Cambal, "Performance Management in Metalworking Processes as a Source of Sustainable Development," in *METAL 2013: 22nd International Conference on Metallurgy and Materials*, Ostrava, 2013, pp. 2017-2022.
- [2] B. Gajdzik, D. Burchart-Korol, "Eco-Innovation in Manufacturing Plants Illustrated with an Example of Steel Products Development," *Metalurgija*, vol. 50, no. 1, pp. 63-66, 2011.
- [3] N. Chamier-Gliszczyński, "Environmental Aspects of Maintenance of Transport Means. End-of Life Stage of Transport Means," *Eksploatacja i Niezawodnos-Maintenance and Reliability*, no. 2, pp. 59-71, 2011.
- [4] R. S. Kaplan, D. P. Norton, "The Balanced Scorecard - Measures That Drive Performance," *Harvard Business Review*, vol. 70, no. 1, pp. 71-79, 1992.
- [5] R. S. Kaplan, D. P. Norton, "Linking the Balanced Scorecard to Strategy," *California Management Review*, vol. 39, no. 1, pp. 53-79, 1996.
- [6] P. R. Niven, "IT Performance Management Using the Balanced Scorecard," in *CIO Best Practices: Enabling Strategic Value with Information Technology*, J. Stenzel, Ed. Hoboken: J. Wiley & Sons, 2007, pp. 185-221.
- [7] N.-G. Olve, A. Sjöstrand, *Balanced Scorecard*. Chichester: J. Wiley & Sons, 2006.
- [8] R. S. Kaplan, D. P. Norton, "Using the Balanced Scorecard as a Strategic Management System," *Harvard Business Review*, vol. 74, no. 1, pp. 75- 85, 1996.
- [9] T. L. Saaty, L. G. Vargas, *Decision Making with the Analytic Network Process: Economic, Political, Social and Technological Applications with Benefits, Opportunities, Costs and Risks*. New York: Springer, 2013.
- [10] Super Decisions: an Introduction [online]. CREATIVE DECISIONS FOUNDATION. Last update 16.03.2013 [cit. 15.3.2013]. Available: <http://www.superdecisions.com/super-decisions-an-introduction/>
- [11] J. B. Butler, S. C. Henderson, C. Raiborn, "Sustainability and the Balanced Scorecard: Integrating Green Measures into Business Reporting," *Management Accounting Quarterly*, vol. 12, no. 2, pp. 1-10, 2011.
- [12] G. P. Kurien, M. N. Qureshi, "Performance Measurement Systems for Green Supply Chains Using Modified Balanced Score Card and Analytical Hierarchical Process," *Scientific Research and Essays*, vol. 7, no. 36, pp. 3149-3161, 2012.
- [13] T. Bieker, C. U. Gminder, *Towards a Sustainability Balanced Scorecard*. St. Gallen: oikos, 2001.
- [14] M. J. Epstein, P. S. Wisner, *Good Neighbors: Implementing Social and Environmental Strategies with the BSC*. Boston: Harvard Business School Publishing, 2001.
- [15] C. W. Hsu, A. H. Hu, C. Y. Chiou, T. C. Chen, "Using the FDM and ANP to Construct a Sustainability Balanced Scorecard for the Semiconductor Industry," *Expert Systems with Applications*, vol. 38, pp. 12891-12899, 2011.
- [16] D. Staš, R. Lenort, P. Wicher, D. Holman, "Conceptual Framework for Assessing the Green Transport Level in Industrial Companies and Supply Chains," *Applied Mechanics and Materials*, vol. 708, pp. 87-92, 2015.

Information Technologies in Logistics Services. Case Study

Izabela Krawczyk-Sokołowska, Katarzyna Łukasik

Abstract— This paper presents a theoretical and empirical considerations about the use of Electronic Freight Exchange in the transport. An essential role in monitoring, coordinating and optimizing the operation of motor vehicles is played by telematics systems, which are used in Electronic Freight Exchanges (EFE). This computer tool facilitates and accelerates logistics services definitely. It is also responsible for creating new business relationships. The main aim of the article is to identify the most important benefits of EFE in logistics and present how does it work in practice.

Keywords—telematics, Electronic Freight Exchange, transport, information exchange

I. INTRODUCTION

The situation on the local and global markets of commercial transactions is very dynamic. More and more frequently, enterprises use IT tools to carry out transactions with contractors in supply chains in the electronic form. The importance of electronic exchange of information via the Internet is steadily increasing, which brings about a reduction in the time of conducting a transaction, global range of operation and an increase in flexibility and reliability of these activities [5]. Due to the application of electronic commerce (e-commerce) in trade, e-business is developing robustly and dynamically. Permanent development of knowledge allows enterprises to use some innovative solutions in the areas of: information technology, telecommunications and production techniques and technologies. As a result of these changes, the integrated management systems of logistics services came into being. The main advantage of the integrated management systems of logistics services is the improvement in the effectiveness of their operation. An essential role in monitoring, coordinating and optimizing the operation of motor vehicles is played by telematics systems, used in Electronic Freight Exchanges (EFE).

II. TELEMATICS IN TRANSPORT

Telematics amounts to telecommunications, IT and information solutions and automatic control methods, which are adapted to the needs of the supported physical systems. Physical systems are installations created as a result of the performance of a specific activity, devices and people operating them, users, and also the environmental conditions, namely, the natural, economic and formal and legal environment. The term “telematics” is most frequently

combined with an adjective specifying the area of its application, e.g. transport, medical, industrial, operational telematics [13].

Telematics in transport consists in providing information and communication using wireless technologies. The main objectives of telematics in transport are [11]:

- to support carriers and distributors in eliminating and reducing delays, too long detours and unplanned downtime,
- to support in avoiding too congested transport routes by directing to alternative roads,
- to reduce the risk of accidents,
- to support travelers by providing information on timetables, connections, changes in courses of public means of transport,
- to increase productivity by reducing (additional and basic) costs,
- to reduce pollutants emitted by vehicles.

The condition to achieve the objectives of transport telematics is constant and regular update of information available and used in the system. In the operation of transport telematics systems, it is essential to promptly respond to changes in weather conditions or traffic congestion. Transport telematics refers to the movement of people and loads using specific means of transport. Telematics services came into being with a view to transport companies but they are also used by clients of these companies. An example is, among others, delivery firms which enable the clients to observe the course of the delivery process of the commissioned consignment on websites.

The frequency and scope of the use of telematics in road transport are increasing steadily. Transport telematics improves transport company management, performance, transport planning and safety, and reduces negative impact on the natural environment. Telematics systems make use of different devices and applications: mobile networks and the Internet, satellite and radio communications, geographical databases, road databases, satellite navigation systems, traffic monitoring equipment (sensors, detectors, cameras, radars), weather monitoring equipment and devices for data transmission to the transport system users [9]. A popular and most frequently applied device in road transport is GPS navigation. The key element of the effectiveness of the navigation is the current update and a very accurate representation of the system of transport infrastructure. However, using only navigation by transport companies

does not allow them to determine the optimal option of the route, which is the source of fuel economy and drivers' working time saving. For transport companies, it is important for the navigation to be an important link of the transport management system in which the shipper, after selecting the route, provides the driver with the information on their choice. The navigation system consists of the three basic functional elements: satellite segment, surveillance and users.

III. EFE – THE CONCEPT AND THE ROLE IN TRANSPORT

Electronic Freight Exchange (EFE) is an information exchange platform between carriers and shipping companies, aimed at facilitating and improving communication and speeding up the conclusion and execution of transactions in the transport industry. An important method to increase savings connected with the operation of transport companies is reducing the presence of empty or incomplete runs of trucks. To improve the effectiveness of the operation of transport companies there are used Internet platforms where there are freight exchanges. The Internet platforms enable the access and exchange of information and entering into transactions, referring to free loads and loading space with the participation of transport, shipping and production companies. Depending on the area and the range of operation, it is possible to distinguish: local, national and international freight exchanges.

In reference to the freight exchange, there can be distinguished two basic elements [1]:

- cargo exchange, which includes the transport offer, that is, goods for carriage,
- vehicle exchange, which contains information on free loading spaces and free runs.

The main task of freight exchanges is to collect orders, share and present the transport offer and information on the free vehicles and to manage the order database from the formal and organizational point of view.

There can be identified two systems of using databases [1]:

- offline freight exchanges – where it is necessary to connect to the database to gain or/and send own offers, then, the connection is ended. The process of introducing own offer takes place before and review of the offers usually after making the connection, when it is already updated. This form of organization of freight exchanges is most frequently used by people who have the Internet access via modem since they bear the cost of the actual connection time, which brings about the possibility of cost reduction.
- online freight exchanges – they consist in transferring own service offers and following the resource of offers already existing in the database of the system, at the time of the Internet connection. This way of organization of freight exchanges is most frequently used by users bearing the costs of Internet connections in the form of the fee, where the fee does not depend on the connection

time. It should be noted that it is a very comfortable and economical way of using the database, which is disposed by the freight exchange.

In Europe, there are more than one hundred different freight exchanges. They are constituted by both small, often free portals and the specialized freight exchanges with a small number of recipients, as well as the European giants in this field, with thousands of users.

“The largest from among the European freight exchanges - Trans.eu. has about 200 thousand users. The next largest freight exchange in the ranking is TimoCom, the services of which are used by most of the German market. An essential supplement to the possibilities of the European freight exchanges is the French exchange - Teleroute. In Europe there are also popular the exchanges from Lithuania (Cargo Lt) and Czech Republic (Raal), but they mostly provide services to the local market [2]”.

On electronic exchanges there are registered shipping, production and transport companies with a diverse resource of cargo trucks. Among cargo trucks there are trucks with a mass of 1.5 tons and tractors with semi-trailer weighing 24 tons, and also custom vehicles, or the ones used for carriage of dangerous goods. On account of a large number of transport companies using the exchange, the shipper does not have a bigger problem with finding specific transport for each type of cargo. Also, the carrier notes an increase in the performance and effectiveness of their transport activity by eliminating or reducing empty carriages, and optimizing the time and routes of the carriage.

The exchange is a huge contact database with the national and foreign companies, which creates an opportunity for the acquisition of standing orders and the establishment of relatively stable, repetitive business relationships with the same contractors. The exchange is an effective tool for the shipper which is an intermediary between two companies: the one possessing the load for carriage and the one providing the carriage. It is also the tool for the shipper which, through the exchange, may find a transport order for a single route and the load for the route back. Carriers also use the exchange very often to find added load, that is, in a situation when they do not have used full loading space. Also, producers who do not dispose their own means of transport and must send their goods to specific destinations are the ones who use the freight exchange. An additional argument, which is significant for manufacturers, is: lower transport costs than shipping costs, which justifies their interest in freight exchanges. Freight exchanges are used both by large production companies and small ones, with a few employees.

IV. THE PRINCIPLES OF EFE OPERATION

Electronic freight exchanges have become a common tool used in the activity of transport and shipping companies. These systems enable posting and the availability of information on loads and free vehicles and loading space. The detailed description of a specific offer and the identification of some relationships are possible due to the appropriate computer application.

“The fundamentals of the construction and operation of electronic exchanges are largely analogous to the construction and operation of other e-business services (e.g. e-stores, auctions). They refer to the documents and ways of communication used in business processes of electronic exchanges. In case of circulation of documents on electronic exchanges, there is assumed: total elimination of documents on paper (e.g. offers, orders, contracts, invoices) and replacing them with electronic documents and electronic acquisition of information, e.g. by electronic forms [...]. In the area of communication on electronic exchanges, the following are aimed at: the elimination of traditional direct meetings (so called “face to face” meetings), organizing only virtual meetings [...], the broadest possible communication automation [...] as well as organizing video-conferences [12]”.

Electronic freight exchanges are a very popular kind of exchange of B2B type, that is, they combine the two parties: shipping companies and transport companies. By means of freight exchanges the company which has a means of transport can find a corresponding offer without a search engine.

The advantages of electronic transport platforms [10]:

- they improve the external communication of the company,
- they make it possible to manage all loads and direct information on the freight to particular recipients,
- they can be treated as platforms for streamlining communication with clients and collecting transport orders,
- they enable access to pan-European freight market,
- they facilitate the effective controlling of all dispatchers working in the system,
- they save working time since it is possible to simultaneously communicate with a number of potential contractors,
- they enable global freight management (the information on all loads is in a single system),
- they contribute to savings in fuel costs and car service,
- they reduce the costs of administrative and office support in shipping companies.

Moreover, freight exchanges improve the performance of transport companies by reducing “empty returns”. Freight exchanges make it possible to find cargo pursuant to the offered loading space, e.g. for deep-frozen products or the carriage of live animals. The advantage of exchanges is also round-the-clock access to offers and their actual current timeliness. A big advantage is also the fact that it is possible to make use of the exchange using a mobile phone.

The basic principle of the operation of cargo exchange and loading spaces is the fact that it operates online. Everyone who disposes of means of transport and does not have a load for carriage, or has a load but they lack free means of transport, may search for contractors on the freight exchange using a computer, a tablet computer or a telephone. Most freight exchanges possess significantly streamlined options of entering information concerning the offer, since an accurate, detailed description of a specific

order is very important as it accelerates and facilitates the successful search for the contractor and entering into the transaction.

To ensure the safety of the course of the service and the settlement of the transaction through online platforms, its supplier must apply appropriate remedial measures. This, most of all, refers to large freight exchanges which cannot afford to unprofessional behavior. “Freight exchanges aim at increasing the safety of their Internet platforms and the transactions conducted there. The companies using the systems available on the market, most of all, are afraid of cargo loss (shipping and production companies) and delays in payment (transport companies). Taking into account the concerns of their clients, the owners of Internet platforms provide appropriate security packages in two areas: transaction security and data security [7]”.

Transaction security is guaranteed under the system of assessment of the company before listing on the exchange, i.e. the analysis of the final situation of the potential participant of the exchange, the date of registration of the business activity and references from the previous contractors. Moreover, freight exchanges introduce the system of assessment of the exchange participants, by means of which it is possible to evaluate the quality and reliability of the load ordering party, which is the participant of the exchange. Some exchanges also possess the debtor reporting system – the company reported as a debtor is listed on the national list of debtors of TSL sector and loses access to the exchange. Freight exchanges, more and more often, have the department of debt recovery and legal advice.

On the other hand, data security is guaranteed on the basis of the personalization of the user, i.e. each participant of the exchange receives a login and password which must be entered before entering the platform and each time while entering into the transaction. Every company which signs up for the exchange is verified by the exchange consultants. As a part of authorization, it must show the documents confirming its credibility, e.g. copies of NIP, REGON, KRS documents and licenses for national and international carriage. Entrepreneurs aiming at the operation on the exchange are verified by the external databases, there are also verified the addresses, phone numbers of the company and the relationships of the client with other companies operating on the market. If the company has lost its financial liquidity or if it is bankrupt, it should not get an opportunity to access the database of the exchange services. Moreover, freight exchanges secure data transfer and access to servers where there is detailed information on each participant of the exchange.

The carrier is obliged to follow the regulations, and the consultants pay a special attention to the attempts of forgery of documents, making accounts available to other companies and spamming the exchange.

Reporting the offer of free vehicles or free loads takes place by filling in a form and providing the basic data concerning the offer, i.e. the place of loading and unloading, type of bodywork or cargo and validity of the offer. When the form is completed, the offer is passed on to the exchange and it is immediately available for shippers, carriers and

logisticians from production and trade companies from the whole Europe. The load or a specific type of a vehicle are searched for in a similar way. While having a computer with the Internet access, it is possible to easily use the system, messenger or route calculator. It is also necessary to install the software. A lot of operators offer a mobile version for a smartphone, which allows to locate the load or vehicle within a few hundred kilometers. The mobile version is used particularly by small shipping companies which, after unloading, urgently wish to find the return load.

Most freight exchanges use the system of users' assessment and comments [2]. It is so called rating system, that is, the carrier, when the order has been executed, may evaluate the reliability of payment and post the comment on the ordering party in their profile. The ordering party also has the right to assess the carrier with reference to the timeliness and quality of the service they provided.

In most systems, there operates the index of reliability of payment of the companies ordering the cargo transport. While calculating the index, there are taken into consideration the financial data of the company and the potential presence on the lists of debtors. The index is most frequently presented in the form of status, i.e. designation as a very good or average payer. Freight exchanges, available on the market, have different systems of charging fees for clients. Some exchanges collect funds for each transaction conducted via the exchange or for a single check of information on the company presenting the offer. However, the most frequently applied is the system where clients pay a single flat-rate fee, which depends on the time for which the subscription is purchased (e.g. monthly, bi-annual, annual). The cost of such subscription usually amounts to PLN 1.5 to 3 thousand per year and depends on the size of the package of additional services [2]. The operators of freight exchanges frequently enable free system testing for a specific time, e.g. a month. It is a very beneficial proposal since it provides, in practice, an opportunity to choose the appropriate fee option, which includes the elements of the system necessary for the transport activity of the user.

Electronic freight exchanges significantly enable the business activity of shipping, production and trade companies. National and international shipping companies find cheap and reliable transport companies on the freight exchange. Producers and trade companies find free means of transport for carriage of goods on the freight exchange and they reduce empty runs of their vehicles. The cargo exchange is a set of offers among which it is possible to come across: free loads, free vehicles, return loads, national and international shipping, transport orders, cheap transport companies, and also transport information.

By means of freight exchanges, companies can reduce costs of transport services and establish new relationships on the market of transport services. To increase the number of contracts and have an opportunity to access a larger number of offers of loads and vehicles, companies often use two cargo exchanges at the same time, which provides them with faster and more efficient transport orders and reduces transport costs.

V. TRANSPORT OPTIMIZATION BY MEANS OF EGT TOOLS

The freight exchange is a tool using the available achievements in the field of techniques and technologies of information transfer and communication. The appearance of new possibilities of communication and information transfer amounts to an opportunity of the implementation of new applications and functionalities for the exchange operators [4]. More and more frequently, it becomes real to currently observe means of transport (e.g. loading). Due to the improvement of the process of making financial clearance via the Internet, also the financial settlement, more and more often, takes an electronic form. Another important area of the exchange operation is the possibility of the automated transfer of information on the phone.

Electronic freight exchanges are a popular type of exchanges of B2B (business to business) type. They bind the carriers, shipping companies and the parties ordering loads. High competition in the sector of TSL services brings about that entrepreneurs are willing to use electronic exchanges which significantly support the management process.

It can be concluded that exchanges operate like a notice board, on the one hand, for shipping companies and carriers and, on the other, the consignors. The main objective of these information and communication systems is the effective exchange of information between the involved parties, improvement in the use of vehicles and reduction in empty runs, which is economically and ecologically significant.

Shipping and transport companies, possessing excess cargo or free spaces in cargo trucks, post their offers on online exchanges. The detailed description of the specific offer and clarification of the specific relationship is possible due to appropriate computer application. The ones interested in loading spaces or loads may select the offer on the basis of regional criteria or properties of goods for carriage. Along with the offer selection on the exchange of transport space, there is presented the information on the load and delivery, type, weight and dimensions of goods, as well as contact and personal details. If the price and the other conditions of the service performance are established, there takes place the immediate transport booking.

Freight exchanges are not really varied in the field of their basic area of operation, their diversity refers to a scope of additional services provided to clients. In Poland, there are several exchanges, among others: *Trans*, *TimoCom Truck&Cargo* and *Wtransnet*.

A useful tool is the possibility of using the freight exchange as the Internet messenger, which makes the information flow between the parties of the commercial transaction more efficient.

Also, the security and credibility of exchange of information and funds on the freight exchange is very important. The *Trans* exchange (which is used for the operation of the investigated company) has the version *Trans 3.0*, including so called *Pakiet Bezpieczna Firma* (the Package of Safe Company), i.e. a set of services increasing financial security of users. This solution is to impede, and finally fully eliminate from the sector, the companies which

should not operate in it. Pakiet Bezpieczna Firma includes the following [8]:

- the system of ratings and comments of contractors – (similar to rating systems of popular auction services), which allows for the assessment of the quality and reliability of the load ordering party,
- TransRisk Index – the indicator of reliability of payment of companies indicated as percentage, and created, among others, on the basis of the data coming from renowned credit information agencies, e.g. Creditreform, Dun&Bradstreet or financial data of the company,
- the system of “Report a debtor!” – as a part of the system there operates the debt exchange, i.e. the list of debtors, containing the data of unreliable companies from TSL sector. Each company, which is on the list, automatically loses access to the exchange.

Clients’ expectations towards the services provided by EGT are varied depending on specialization, market or size of the company. However, the standard required by all the participants is an accurate, actual and currently monitored map. “Modern software offered to transport companies should be possibly intuitive and easy to operate (and even maintenance-free). In a multitude of different duties, the client will choose such solutions which will not disrupt the course of their classical work process but they will just bring about saving time and funds [6].”

VI. CASE STUDY. THE ANALYSIS OF THE USE OF THE ELECTRONIC FREIGHT EXCHANGE IN THE INVESTIGATED COMPANY

X-Car is a shipping and transport company which operates in most countries of the European Union. The company operates twenty four hours a day, seven days a week. The greatest asset of the enterprise is the rapidity of operation since it provides the delivery of loads within 24 hours to most European cities. If there is such a necessity there is sent a double manned vehicle, which influences the speed of the cargo delivery even at large distances (JIT: Just in time, ASAP: as soon as possible). X-Car cooperates with the largest shippers in Europe and the leading companies from the automotive industry.

In its activity, X-Car uses the Trans exchange, i.e. the platform of exchange of information on free loads and trucks from the whole Europe. The Trans exchange disposes one of the largest databases of transport offers in Europe.

The Trans exchange provides a range of benefits for trade and production companies, shippers and carriers. The benefits for shipping companies, resulting from the use of the exchange, amount to: up to 150 000 offers of loads and vehicles daily, efficient and fast execution of transport orders and cooperation with certified carriers. For carriers, the most important benefits are: prompt establishment of cooperation by a messenger, the system of assessment and reliability of payment of contractors, and also a new return load every 0.9 second. Trade and production companies, due to the Trans exchange, may easily establish the cooperation by a messenger, efficiently and timely arrange the transport of loads and are supplied with efficient security procedures [3].

The Trans system is a package of solutions which enable the establishment and implementation of the efficient and secure collaboration: carriers, shippers and the ones carrying loads in the whole Europe. The Trans system consists of the following solutions[3]:

- a simple system of searching for and posting offers,
- monitoring the most interesting offers, the possibility of filtering by the type and kind of the offer, the required bodywork, physical parameters of load, properties of the freight or the assessment of the carrier;
- a simple system of posting offers in the form of intuitive application form, facilitating the specification of the detailed parameters of the load;
- text messaging and voice communication with the contractor;
- detailed information on the contractor, i.e. address details of the company, the description of its activity, and also the required registration documents and licenses.

The analyzed X-Car company uses the Trans freight exchange because they have confidence in it. The exchange provides the security of transactions, makes it possible to check the documents of the carrier and to find out the assessment and opinions of other users. Ratings and opinions posted by users are an important source of information – particularly, if the company wants to establish the cooperation for the first time. There is a possibility to rate the specific company in the three-tier system, that is positive, neutral and negative evaluation. Ratings and comments of clients are, therefore, a tool allowing the exchange users to assess their contractors: the carriers – to assess the solvency and credibility of the ordering parties, and the companies ordering the carriage – to express their opinion on the professionalism, timeliness or quality of the service provision by sub-contractors. The information on the assessment received by the company is posted next to its offers on the exchange, and also in a bookmark “Information on the Company”.

The Trans exchange uses TransRisk Index, i.e. the indicator of the reliability of payment for the companies ordering loads. This indicator is calculated by a special algorithm and it contains important information for carriers whether it is worth trying to cooperate with a specific company ordering load or if a potential contractor is a reliable payer. The analyzed X-Car company, in its activity, uses not only the services of the Trans freight exchange. As a tool supporting the transport offered by the company there are also used the ViaMichelin maps of Europe. It helps calculate and plan the routes taken by vehicles.

While analyzing the activity of the X-Car transport company, it should be noted that it provides complex transport and shipping services in the field of carriage of loads in the area of the European Union. The activity of the X-Car company, without using supporting tools, that is, electronic platform, would be significantly limited and less efficient. Due to the Trans exchange, X-Car cooperates with many transport companies, using the access to the global network of information and contacts. The X-Car company

must run its activity on the market of transport services on the basis of the electronic freight exchange, but also on the basis of accurate and currently updated maps of Europe. Using the whole of the available information and supporting tools by the company will allow X-Car to accept orders, determine the optimal conditions of carriage, drivers' working time, and also the price level of the provided transport services.

VII. CONCLUSIONS

The use of the freight exchange allows transport companies to reduce empty carriage, optimally use loading space, added load and to gain additional transport orders and, consequently, to reduce costs of providing services and gain additional profit.

The main advantage of electronic transport platforms is efficient external communication of the company. The analyzed X-Car company, in its activity, uses the Trans exchange, due to which it easily finds orders for carriage of goods, and also reduces empty carriage. The exchange also provides the detailed information on every user, along with the assessment of their activity and payments made on time.

It should be noted that not every company may become the user of the Trans exchange. The exchange provides security through a perfectly working system of monitoring the companies, which intend to sign up and become the exchange users.

The rapid rise in popularity of online exchanges mainly results from reducing the operation costs of companies and the ease of acquisition of business contractors and the course of completion of new transactions. The attractiveness and availability of electronic platforms for different companies is connected with the openness of the Internet and low costs of using electronic platforms. In the current economic relations, long-term, fixed contracts are abandoned for the benefit of the current and ad-hoc search for business partners who offer the most favorable conditions of cooperation.

REFERENCES

- [1] T. Grzelak, „Zastosowanie technologii GPS, GPRS, wykorzystanie Internetu oraz systemu zdalnego zarządzania ruchem (ATMS) we współczesnym transporcie”, [in:] Współczesne procesy i zjawiska w transporcie, Uniwersytet Szczeciński, Szczecin 2006, p. 31, at: http://www.wzieu.pl/zn/447/ZN_447.pdf - accessed on 12.02.2015
- [2] <http://m.forsal.pl/branze/finanse/w-transporcie-drogowym-bez-gieldy-ani-rusz> - accessed on 10.03.2015
- [3] <http://www.trans.eu/pl/> - accessed on 5.02.2014
- [4] M. Jurczak, „Gielda kontra gielda”, Transport i spedycja No 4/2011, p. 29.
- [5] I. Krawczyk- Sokółowska, B. Ziółkowska, “Computer- Aided and Web- based tools in customer relationship management”, Acta Electrotechnica et Informatica. Vol.13, No.4, 2013, p.13.
- [6] M. Klecha, “Bezstratne planowanie”, Top Logistyk, Nr 1/2014 – accessed on 5.05.2014 at: <http://logistyczny.com/artukul.php?id=5563>
- [7] J. Lewandowska, „Gieldy transportowe”, Wyższa Szkoła Logistyki w Poznaniu, accessed on 1.09.2012, Opracowanie prezentacji współfinansowane przez Unię Europejską w ramach Europejskiego Funduszu Społecznego
file:///C:/Users/dom/Downloads/Gie%C5%82dy%20transportowe.pdf
- [8] M. Loos, “Ratunek na brak frachtów”, accessed on 5.05.2014 at: <http://www.log24.pl/artykuly/ratunek-na-brak-frachtow,22>
- [9] J. Mikulski, I. Nowak, „Telematyka – przyszłość transportu i logistyki”, Logistyka 2/2010.
- [10] P. Romanow, „Internetowe giełdy frachtów w operacyjnej działalności przewoźników drogowych”, Logistyka 5/2011 – accessed on 02.04.2014 at: http://www.e-fakty.pl/index.php?option=com_content&task=view&id=6060&Itemid=73
- [11] R. Sałek, M. Kliś, „Zastosowanie systemów telematycznych w zarządzaniu przedsiębiorstwem transportowym”, [in:] “Teoretyczne i praktyczne aspekty zarządzania przedsiębiorstwem”, Sekcja Wydawnictwo Wydziału Zarządzania Politechniki Częstochowskiej, Częstochowa 2012, p.70.
- [12] W. Wiczerzycki, „Giełdy elektroniczne”, [in:] Instrumenty zarządzania łańcuchami dostaw, M. Ciesielski [ed.], Polskie Wydawnictwo Ekonomiczne, Warszawa 2009, p. 331.
- [13] K.B. Wydro, „Telematyka – znaczenie i definicja terminu”, Telekomunikacja i Techniki Informacyjne, 1-2/2005, Instytut Łączności. Państwowy Instytut Badawczy, Warszawa 2005, p. 117.

First Author: **Ph.D. Izabela Krawczyk- Sokółowska**, an Assistant Professor at the Faculty of Management in the Czestochowa University of Technology. She is the author of numerous publications in the field of innovativeness of companies. The most important publications: 1) main autor I. Krawczyk- Sokółowska, *Analysis of Profile of Financing Innovations in an Enterprise*. Transport & Logistics. Carpathian Logistics Congress. 27. - 30. September. 2011, Podbanske, High Tatras, Slovak Republic.2011, 2) main autor I. Krawczyk- Sokółowska, co- autor B. Ziółkowska, *Computer-Assisted Tools in Customer Relationship Management*. Department of Computers and Informatics, INFORMATICS 2013. FEEI TU of Kosice. Proceedings of the Twelfth International Conference on Informatics. Eds. Valerie Novitzka, Stefan Hudak. November. 2013, 3)main autor I. Krawczyk- Sokółowska, *The Role of Human Resources in an Innovative Enterprise*. The Publications of the MultiScience - XXVIII. microCAD International Multidisciplinary Scientific Conference. Miskolc, 10-11 April.2014; E-mail address: sokoliza1@o2.pl

The second Author: **Ph.D. Katarzyna Łukasik** an Assistant Professor at the Faculty of Management in the Czestochowa University of Technology. She is the author of numerous publications in the field of organizational culture and modern concepts of management. The most important publications: 1) main author K. Łukasik, co-author K. Brendzel-Skowera, H. Kościelniak, *The Impact of Organizational Culture on Knowledge Management in the Light of Empirical Studies*, 14th EBES Conference - Barcelona. Proceeding CD. October 23-25, 2014, Barcelona, Spain. Vol.1; 2) main author K. Łukasik, co-author B. Ziółkowska, *Occupational and Geographical Mobility of Intellectual Capital*, 15th EBES Conference - Lisbon. January 8-10, 2015, Lisbon, Portugal. Proceeding CD. Vol.2; 3) main author K. Łukasik, co-author K. Brendzel-Skowera, *The Research on Entrepreneurial Organizational Culture, Management and Managers Facing Challenges of the 21st Century*. Theoretical Background and Practical Applications. Monograph. Eds. Felicjan Byłok, Iveta Ubreziowa, Leszek Cichobłaziński, Szent Istvan Egyetemi Kiado Nonprofit Kft., Godollo 2014; E-mail address: kasia2lukasik@op.pl

The Development of e-Business Services in Poland

Elzbieta Wyslocka and Renata Biadacz

Abstract—Services based on the potential of information technology are named e-services. They are particularly noticeable in the industries related to finance and consulting. In this context no doubt that they have critical significance for a sector of modern business services. Today's business services primarily consist of centers with foreign capital, such as a common service center, business processes outsourcing, IT outsourcing as well as research and development (R & D) center.

The article presents considerations whose goal is to underline the importance of e-business process outsourcing services, including financial-accounting services in today's world, when we stand in front of a vast variety of social and technological challenges. The purpose of this work is to obtain knowledge of utilization and development of financial services in outsourcing as well as e-business services in Poland.

The research is based on the need to deploy solutions in order to improve business services. The methodical base to this dissertation consists of the literature analysis and reports developed by the Leaders of Business Services containing data for development of e-business services in Poland.

Keywords—accounting, services, e-business.

I. INTRODUCTION

Nowadays the Internet has become a medium, which determines almost every sphere of human activity. It has an impact on conducting the business, settlement of official errands, leisure time, and even shaping human relationships. From a business perspective, it is noticeable that the Internet facilitates creation of new business areas, which use virtual marketplace. Through the Internet, companies can offer a lot of services, often unprecedented in the real world [1]. This is particularly evident in financial sectors (insurance, banking, accounting), and related to widely understood consulting. This approach of companies and organizations is to develop a new model of provision of services based on the potential of information technology [2], so-called e-services.

There are many definitions of e-services in the literature. Because of difficulties in defining e-service, it is usually recognized and understood as a service that meets the following conditions: - is provided in partially or fully automated way by information technology, - is carried out on the Internet and via the Internet, - is individualized with

respect to the recipient (personalized), - parties of the service are located in different locations (remote service) [3]. According to another, much broader approach, e-services are understood as a new form of providing services through the Internet, from the moment of contact of the company with the client in order to present the offer, through ordering services, providing it and contacting with the client after the end of contract [4, 5].

E-services market, both in Poland and in the world is growing rapidly, covering all sectors and all areas of the traditional market. Many Internet services occur because of the simplicity and greater, as opposed to the traditional market, availability to the customer. Relatively low start-up cost and a very large potential market audience are also important elements [2]. The growing popularity of the Internet and at the same time easier access to it cause that there are no restrictions in this type of communication in the form of language barrier, the time or location [6].

Research and megatrends indicate that the most rapidly developing areas of business, in which e-services will play a key role, due to the need to reduce costs and save time, include [1]:

- support of business (professional business services, financial intermediation),
- electronic banking services,
- cloud computing and web solutions,
- e-commerce, - education and training of employees (e-learning) [7].

It should also be noted that modern enterprises, aiming to run business more efficiently must be constantly active, which is expressed, among others, in the ability to continuously develop and adopt new ideas and create new value. Often, it is emphasized that presently, only those companies that are flexible enough and can adapt to demands of an increasingly dynamic and globalized environment, will be able to survive and develop [8]. The complexity of the business, the level of complexity of the manufacturing processes, relentless pressure to reduce costs and increasingly common trend among companies to focus on key activities (*core business*) make some or all of the processes (activities, resources) increasingly entrusted to specialized companies. Using outsourcing becomes more and more important, which largely reflects the changes taking place in the business environment of enterprises [9]. Although the concept of outsourcing is not a new concept, it is the process, which systematically gains

E. Wyslocka (corresponding autor) and R. Biadacz are with the Department of Management, Czestochowa University of Technology, 42-200 Czestochowa, Poland (phone: +48 601-209-175; fax: +48 34-361-38-76; e-mail: wyslocka@zim.pcz.pl).

greater popularity, as indicated by studies carried out in 2011 by the Outsourcing Institute [10].

M. Trotsky defines outsourcing as "a project for a spin-off of realized functions from the organizational structure of the parent company and transferring them to external service providers" [11].

W. M. Lankford and F. Parsa indicate that outsourcing involves disposing of tasks, which may be waived or that can be done cheaper and often better outside the company. The company focuses on their core competencies, which represent competitive advantage. In contrast, areas that don't provide this advantage, and are often only a secondary or incidental, are separated from the business unit processes and carried out by external companies [12]. In this approach, the authors draw attention to new features on this phenomenon, such as the cost and quality of external services. Both factors play a significant role. The use of outsourcing is in fact justified when the outsourced functions are performed better or at least not differ from the current level. Costs should be lower than those generated by the business prior to the separation of a particular function [8].

In the broadest perspective outsourcing is seen as a restructuring project, which aims to increase the flexibility of actions by reducing the unit's organizational structure through reduction in the number of cells, organizational positions and management levels, which lead to a reduction of employment and thus reduction of costs. It is connected with the management concept of Lean Management [14].

Among types of outsourcing the following concepts can be distinguished: offshoring (moving business to other countries), nearshoring (ordering part of the function to the same or a neighboring country to be executed by an external firm) and centralized services, involving separation of certain activities in several divisions of the company and moving them to one center owned or controlled by the parent company [15].

The concept of outsourcing is also linked with such terms as subcontracting (which is a transfer of part of the burden of the contract to another, independent business entity, while the order is made by the main contractor for the project and must assume responsibility for the execution of the whole project), e-outsourcing (is to control the entire outsourced process of with the help of e-business solutions, using the Internet) [9].

Outsourcing has gained particular importance in the 80s and 90s, when companies began to abandon their non-core activities. Initially, it covered mainly the huge data centers and spread to other areas of information technology, e.g. network management and storage management. In a further stage, it was related to administrative processes within the enterprise. At that time, the outsourcing of payroll, staff consulting, human resource management, and outsourcing of financial and accounting processes have been established in large enterprises. The next stage in the development of outsourcing is to isolate the key operational processes in the enterprise - logistics outsourcing, call centers, operations of the banking facilities.

II. OUTSOURCING OF FINANCIAL AND ACCOUNTING SERVICES

On the basis of aforementioned considerations, it can be concluded that outsourcing is a kind of long-term cooperation and the special bond between organizations. A typical services outsourced to an external entity are financial and accounting services. Outsourcing of financial and accounting management as a strategy is important for companies in order to maintain a growth of its effectiveness. This is a relatively new concept in the management of the company, which is becoming more and more popular [13]. Of course, small companies often use the services of accounting offices and abandon running bookkeeping inside the unit. Note, however, that larger companies can also benefit from the financial and accounting outsourcing. Along with the growing scale of business, the complexity of performed tasks increases, and thus there is more possibility of using outsourcing and the potential benefits associated with this decision are greater. It is worth noting that not only accounting services are in high demand among entrepreneurs, but also payroll or financial services experience increased interest in outsourcing. The result is that outsourcing companies expand the package of provided services by these much-related services. In order to improve and broaden the scope of their services, they employ highly qualified specialists (HR, payroll, accounting), but also more and more often, due to the considerable complexity of the problems, collaborate with specialized tax offices. This allows to maintain the high quality of services, and solve problems in a broad sense.

The main advantages associated with a decision to outsource, especially accounting, include [16]:

- possibility of cost savings,
- greater financial flexibility resulting from the reduction of fixed costs,
- desire or need to focus on key activities,
- flexibility in employment,
- improvement in the quality of implementation of the accounting function.

It should also be noted that the accounting service provider, i.e. for example accounting offices, thanks to dozens of clients achieve significant economies of scale, which are not available for small and medium-sized enterprises, implementing the accounting function within their organizational structures. It represents also a significant reduction in the risk, because it is the accounting office, which at this point takes over the risk and is responsible for any errors in the accounts. A relatively high liability insurance policy held by the service provided contributes to the safety of clients.

Another important aspect of risk reduction that occurs when using financial and accounting services outsourcing is protection against the negative consequences of unexpected absences (e.g. disease) of staff performing the function of chief accountant. Provider of accounting services, or outsourcing companies, has many professionals who can replace each other, and the disease of any of them does not affect the price of the service, or interfere with the course of its implementation.

Transaction processes, which seem to be simple, repetitive, and mass, are frequently subject of financial and accounting outsourcing in the organization. This translates into the release of specialized employees from the performance of standard, repetitive tasks. It allows using their potential in the processes, the execution of which is increasingly expected in financial and accounting departments - the processes responsible for the processing and delivery of information to support decision-making, often shaping the competitive advantage of the organization. What is more, expectations towards such processes intensively grow - information must be generated faster and must be more accurate. They should now be the main activity of the finance and accounting departments [17].

III. THE DEVELOPMENT OF THE MODERN BUSINESS SERVICES SECTOR IN POLAND

The development of the business services sector is one of the most important manifestations of globalization. It is a source of enterprise competitive advantage and determines the possibilities of their development. Poland is one of the most interesting locations on the map of global business services. The data indicate that more than 10 years ago, there was no such business services sector in Poland, and over the years the development was so strong, that Poland became number 1 in Europe and number 3 in the world [18].

In Poland there are already 470 service centers with foreign capital, belonging to 325 investors and employing 128 000 people (as of 30.04.2014). 66 new service centers have been established since the beginning of 2013. New investors, which did not have service centers in Poland yet, have formed 60% of them.

The ten largest business service centers (Cracow, Warsaw, Wroclaw, Gdansk Agglomeration, Lodz, Katowice Agglomeration, Poznan, Bydgoszcz, Szczecin, Lublin) employed a total of 95% of all employees service centers with foreign capital in Poland.

Business services sector is therefore one of the fastest growing sectors of the Polish economy. Financial sector in Poland, which grows rapidly, is a forge of human resources for the business service centers. The financial services sector (*BIFS - banking, insurance, financial services*) are activities for banks, insurers and other financial institutions. Polish financial services sector is the biggest in Central and Eastern Europe. For several years, is also one of the fastest growing markets in Europe. In the field of business services, a clear trend has been noticeable for several years, shifting the focus from "traditional" customer service to solutions based on specialized knowledge (*Knowledge Process Outsourcing*) and highly qualified staff. This applies to both the core processes of the organization, as well as supporting functions such as finance, accounting, HR and IT. Understanding market segment (i.e. *vertical*) for which the services are performed, and access to experts, is the key to success. After a slowdown that took place in 2008-2009, in the years 2010 and 2011, the

global market for outsourcing services for the financial sector has been growing at more than 16-percent annual rate [19].

Processes shaping the international division of labor in the global economy led to the strengthening of the Polish position as one of the most often chosen service centers location for business process outsourcing (BPO, ITO, SSC, R&D) by the investor¹.

Employment in the modern business services sector in Poland has been steadily increasing. Compared to the beginning of 2012 the number of employees of service centers financed by foreign capital increased by over 50% - from 83 000 to 128 000 people. Over the past 12 months (from April 2013) at least 18 000 new jobs have been established, of which the most in Cracow (> 5 000). The average annual increase in employment in the country in a few years (from 2009) stood at about 15 000 people. Centers in Poland have 10 out of the 25 largest banks in the world (Deutsche Bank, HSBC, BNP Paribas, RBS, Citigroup, ING, Santander, UBS, UniCredit and Credit Suisse), now making Poland one of the Europe's leading BIFS business services sector. It is also worth to note a number of Polish outsourcing companies working for clients in the financial sector. Examples include among others: Ericpol, Outsourcing Experts Group (OEX), Casus Finance. It is worth to underline that the vast majority of service centers with foreign capital in Poland, acting on behalf of the financial industry, delivers offshored services, working mainly for their counterparts from Western Europe and the USA. It is estimated that services, which came to Poland thanks to offshoring, are provided by several thousand people - 2/3 of center employees working with foreign capital for the financial industry (i.e. also outsourcing companies) [19].

Given the current development of the industry, it can be concluded that centers located on Polish territory, will be able to carry out all operations for investment banks, international brokerage, custodian banks and other financial institutions in the near future. It should be noted that Polish workers are well prepared to provide services to entities in the financial sector. Knowledge of foreign legal conditions (including regulatory issues) related to the implementation of business processes along with the ability to provide services in several world languages, represent an added value for service centers located in Poland.

There are several organizational models of service activities in BIFS sector. In case of financial services the most common type are hybrid models, in which the company uses both outsourcing and services offered by captive centers. The second, slightly less popular operating model for BIFS services are shared service centers (*captive centers*). It is worth noting that the financial services sector has been one of the first, in which such operating centers were established. The

¹ The study adopted a broad definition of the industry, including such activities as: shared service centers (SSC), companies providing business process outsourcing (BPO) and IT outsourcing (ITO) as well as centers of research and development (R&D). This definition was adopted after the report Sektor nowoczesnych usług biznesowych w Polsce, Association of Business Service Leaders In Poland (ABSL), 2014, p.7

popularity of shared services on a global scale is not decreasing. The activities performed by these centers include both supporting key processes (*core processes*, e.g. for BIFS sector companies these are risk analysis, key processes related to capital markets, credit card services, etc.) and support services (*support*, for example IT and HR support). On the other hand, BPO centers (*business process outsourcing*) usually offer support services, but in Poland more often have the ability to support key business processes [19]. Services in the BIFS sector experience a rapid development due to the need for cost optimization, standardization of processes and the development of technology, the need to comply with regulatory requirements and to obtain access to talented employees.

In recent years, a financial and accounting processes in Poland were provided by the largest number of business service centers. It should be noted that the centers usually offer at least two such processes. Finance and accounting and IT services. Centers operating in Poland won and still maintain many major global customers. A common challenge undertaken by centers operating in Poland is the need for standardization of financial and accounting processes between local client, enabling the client transformation processes at the global level and supporting the expansion plans of the client. Often, provision of financial and accounting services requires to meet the customer expectations concerning not only reduction of operating costs, but also improvement of the financial indicators, including on recovery of or collection of overdue receivables and the time to deal with potential disputes. At the same time, such actions must be accompanied by a high level of service to maintain key customers, posing problems with timeliness, but subject to custom service [19].

An example might be Capgemini², which, among others has been providing financial and accounting services to one of the world's market leaders in insurance products and services for 8 years. This required the transfer of operations from the client to the local Polish offices, which resulted in the transfer of knowledge and documenting processes by describing them in the form of hundreds of detailed procedures. Currently, this client service is being provided by a fully dedicated workers in two Capgemini locations: more than 320 people work in the center of Cracow in Poland, and more than 130 people in the center in Chennai, India. Polish team specializes in providing the most advanced and demanding language skills (i.e. *Value Centre*), while workers in India are focused on standard operation with a large volume of transactions (i.e. *Transactional Centre*) [19]. Range of financial and accounting services in Poland can be very broad and include (as in the case of Capgemini) among others: invoicing, accounting for transactions in the books and reconciliation of bank accounts,

² Capgemini Polska belongs to the Capgemini group and employs more than 5,000 professionals, providing services for the telecommunication, banking, insurance industries and utilities. The Kraków company provides services in 29 languages. Apart of a wide range of business services, Capgemini Poland also implement information systems, integrate IT services and solutions in the field of IT infrastructure.

fixed assets accounting, receivables management and debt management, customer relationships management, settlements between subsidiaries, working with client entities to resolve disputes related to client invoices and documents, reinsurance accounting, closing settlement periods on ledger accounts, reporting and regulated reporting in the home countries [19].

IV. SERVICE CENTERS (BPO, ITO, SSC, R & D) WITH POLISH CAPITAL

Business service centers with Polish capital have very stable prospects. Growing demand for outsourcing services in Poland has among others a great impact on it. Despite the fact that the vast majority of enterprises in Poland use outsourcing in many areas of activity, it is still used in a relatively limited extent. Moreover, the full potential arising from the use of shared services center model is still not fully used. Group of several dozen leading centers of Polish capital - outsourcing (BPO, ITO), shared services (SSC), and research and development (R&D) - employs more than 50,000 people. Depending on the adopted criteria for selecting companies in terms of business profile and size of the business, particularly in the context of outsourcing business, this market can be considered as significantly wider. Taking into account, inter alia, numerous internal customer service centers and all outsourcing service activities (not necessarily done in typical service centers), the size of the domestic services sector for business or employment in firms with Polish capital, can now be estimated at more than 200,000 employees [20]. Scale of operations of Polish outsourcing centers is evidenced by range of supported markets. More than 81% of domestic outsourcing companies apart from Polish companies also provide services to foreign clients. Diversification of markets is beneficial for companies because is conducive to the development and stabilize their operations. Among the companies offering financial and accounting outsourcing services are Target BPO Polska (for five years has been offering its accounting services also for the Scandinavian market) [21] and Business Support Solutions (BSS) belonging to the Pelion Healthcare group (the company by subtracting cooperation with foreign customers, including Belgium and the Netherlands develops a model of international outsourcing) [22].

Beside the above-described selected Polish outsourcing companies and R&D centers, there are also a number of shared service centers. Among the largest are centers owned by large companies, where the State Treasury is a significant shareholder, including PZU, Poczta Polska, Tauron, Katowicki Holding Weglowy, PGNiG (PGNiG Service).

A growing number of investors do not perceive Poland as a whole, but concentrates on the various centers in the country. At the same time, investors are not only interested in major Polish cities, but also in the "emerging" centers, such as Bydgoszcz, Lublin, Radom and Szczecin. Compared with other countries in Central and Eastern Europe, Poland has as many as 11 cities with a population of more than 300,000, as

well as smaller with excellent conditions for the development of business services.

V. CZESTOCHOWA - A NEW LOCATION ON THE MAP OF BUSINESS SERVICES

Due to the rapid development of the business services sector in major Polish cities, there is increasing competition among employers and the labor market becomes gradually saturated. For this reason, the preferred option for many companies may be diversification of places of business, for example, by creating a support unit or branch office in another smaller academic city, which is well-communicated with the main center. The undeniable advantage of Czestochowa is location near one of the leading centers of the sector in Poland - Katowice Agglomeration. Significantly lower costs of renting office space in comparison to larger centers in Poland is an important advantage of Czestochowa. The office space market in the city will be growing rapidly, which responds to the needs of potential investors in the business services sector. Less competition between employers, lowers employee turnover and builds stronger relationships between workers and their employers. Czestochowa positions itself among well-communicated academic centers characterized by the availability of an educated potential employees and the development of office infrastructure, which allows to predict that the city will be in the next few years on the lists of potential locations for new investments in BPO/ITO/SSC service centers more often [23]. The e-business services currently operating in Czestochowa are shown in Table 1.

One of the fastest growing companies on the Czestochowa market is TRW Automotive. The US company TRW Automotive is a global leader in automotive safety systems and one of the largest in terms of sales value of companies in the automotive industry. TRW currently employs over 60 000 employees in 185 locations in every region of the world where cars are manufactured. In Czestochowa has both European Shared Services Center and Engineering Center.

European Shared Services Center (ESSC) TRW in Czestochowa was founded in late 2006 in order to manage key financial and accounting operations of the European plants of TRW. It provides services by standardizing processes, complete service, the use of best practices, economies of scale, and the use of modern information technologies. Currently employs over 230 people, serving more than 40 units of TRW in European countries such as Germany, Great Britain, France, Poland, Czech Republic, Spain, Portugal, Italy, Slovakia, Romania. In connection with the development, ESSC currently plans to hire more people to collaborate on new projects related to handling accounting processes.

VI. CONCLUSION

Polish BIFS sector (banking, insurance, financial services) is one of the leading markets of this type in Central and Eastern Europe. The level of sophistication of supported processes has increased rapidly in the Polish sector of modern

business services the number of leading global brands has grown. The following trends can be observed in the development of this market:

- Developing business services market in Poland becomes very attractive for BPO entities of all sizes and various countries of origin - these are, for example, Indian companies, as well as investors from the Middle East and Africa, who consider Poland as a potential location for their business services or as a market for potential mergers and acquisitions. Many SSC/BPO centers operating in Poland - with varying levels of maturity, size and scope of provided services - make an excellent opportunity for investors to enter the European BPO market.

- More and more smaller companies set up their business services centers in Poland, while so far it was a practice generally exploited by larger entities. For many companies, the creation of a service center or other SSC operations in Poland often allows for more flexibility compared to some countries in Western Europe. Moreover, the "brand", quality, effort and innovation of Polish workers are seen as a major advantage.

- Recognized service centers develop their services towards more advanced processes, moving up in the "value chain" and adding knowledge-based processes, as well as introducing new functions within the scope of their activities (for example, marketing, supply chain, legal services) to provide multi-functional business support.

Key factors for investors who have decided to locate their services in Poland are still the cost and availability of the "right" talent pool [20]. However over the past two years, two specific elements on this location has started to become more and more important: how various organizations in the city/area work together, and how business cooperation with universities and within the business services sector looks like. Additionally, the extent to which local government supports the business services sector became significant. It matters not only at the time when investors are attracted and initial cooperation begins, but also in terms of regular activity and in the growth phase.

Looking to the future of the business services market, we see that as a result of the global economic crisis, some Western countries - such as Spain, Portugal and Italy - become more frequently seen in a long list of a potential locations. Poland has a strong advantage in terms of a number of other important criteria. The overall outlook for the business services sector in Poland is very positive. Poland is and will continue to be seen as a mature market for business services, attracting new investors. There will be companies opening their SSC, BPO/ITO centers extending its reach and new operators seeking to take over the existing service centers in Poland. As in previous years, the service centers will continue to grow, attracting new geographic areas, moving from performing one function to multi-functional mode, as well as introducing more advanced activities within services already provided. Given that a critical mass of business services industry in the country has already been achieved, it seems that no location in Central

Table 1

Selected investors present in Czestochowa

Name	Country	Business profile
TRW Automotive - TRW Polska Sp. z o.o. Financial Centre Engineering Centre	USA	TRW Financial Services Centre (European Shared Services Center - ESSC);
		Engineering Centre - specializing in the design of automotive safety systems;
LGBS	Poland	IT outsourcing; IT services including software design and development, consulting and IT consulting;
TeleConcept, (Loyd Capital Group), Contact Center One, Telbridge, Polcall, Call Center Inter Galactica	Poland	business services in the field of customer service

Source: prepared by ABSL based on information from companies and their websites

and Eastern Europe will not be able to compete in the coming years with Poland. Polish cities will continue to strengthen its position on the world map of business services. Growing number of them will also appear in the general rankings and reports on local research. At the same time, renewed efforts and cooperation must be continued to ensure Poland a reputation of the country in which you can easily invest in more advanced processes and functions. The picture of the country where cooperation with the public administration, education and business sector facilitates access in the market and improvement in the general conditions enables further development of the business services industry must be maintained.

REFERENCES

- [1] K. Batko, G. Billewicz, "E-usługi w biznesie i administracji publicznej". Available: http://www.ue.katowice.pl/uploads/media/3_K.Batko_G.Billewicz_E-uslugi_w_biznesie....pdf.
- [2] W. Gryncewicz, K. Lopacinski, "Innowacyjna koncepcja świadczenia usług z wykorzystaniem zintegrowanych rozwiązań informatycznych w obszarze zdrowia", in: *Technologie informacyjne w kreowaniu przedsiębiorczości*, A. Nowicki, D. Jelonek, Ed. Czestochowa: Wydawnictwo WZ Politechniki Czestochowskiej, 2014.
- [3] M. Sliwinski, *Modele biznesowe e-usług*, Warsaw: PARP, 2008. Available: http://www.web.gov.pl/g2/big/2009_03/9f8f4a02eb05becf56a9f7320c00390f.pdf, read on 25.02.2015.
- [4] A. Dabrowska, M. Janos-Kreso, A. Wodkowski: *E-usługi a społeczeństwo informacyjne*. Warsaw: Difin, 2009, p. 41.
- [5] D. Cotirlea, "Issues regarding e-service quality Management - customization on online tourism domain", *Polish Journal of Management Studies*, vol. 3, p. 33-34, 2011.
- [6] D. Wielgorka, J. Szymczykiewicz, "Chmura obliczeniowa jako nowoczesne rozwiązanie usług IT dla przedsiębiorstw", in *Współczesne problemy zarządzania w podmiotach gospodarczych i publicznych*. D. Wielgórka, Ed. Czestochowa: Wyd. WZ, 2014.
- [7] *Spółeczeństwo informacyjne w liczbach*. V. Szymanek Ed. Departament Społeczeństwa Informacyjnego, Warsaw 2012.
- [8] T. Kopczyński, *Outsourcing w zarządzaniu przedsiębiorstwami*, Warsaw: PWE, 2010, pp. 7-8.
- [9] J. Grabowska, "Outsourcing usług logistycznych", *Scientific Journals of Silesian University of Technology*, series: Organizacja i Zarządzanie, 60, Gliwice, 2012, pp. 83-84.
- [10] *Trendy w outsourcingu w Polsce*. Elaboration of Outsourcing Institute, November 2011.
- [11] M. Trocki, "Outsourcing. Metoda restrukturyzacji działalności gospodarczej", in *Outsourcing logistyczny*. Logistyka, no 6, A. Jonkisz, J. Jaroszyński, Instytut Logistyki i Magazynowania, 2008.
- [12] W. M. Lankford, F. Parsa, "Outsourcing: A Primer", *Management Decision*, 1999, 37/4.
- [13] C. Pop Sitar, "Optimization of Management decisions for purchasing of business services", *Polish Journal of Management Studies*, vol. 5, pp. 213-215, 2012.
- [14] B. Nadolna, "Outsourcing", in *Od audytu do sponsoringu w rachunkowości*, K. Czubakowska, Ed. Warsaw: PWE, 2007, p.200.
- [15] E. Wyslocka, "Outsourcing usług księgowych w badaniach czestochowskich biur rachunkowych", *Scientific Journals of the Szczecin University*, no. 668., pp. 331-340.
- [16] *Biznes. Zarządzanie firma*, Tom 1, Warsaw: Wydawnictwo Naukowe PWN, 2007, p. 218.
- [17] M. Kawa, "Jaki jest potencjał outsourcingu finansowo – księgowego?" *Controlling i Zarządzanie*, no 1/2015, p.39.
- [18] "Polska wśród liderów branży usług dla sektora finansowego". Available: <http://biznes.onet.pl/wiadomosci/kraj/polska-wsrod-liderow-branzы-uslug-dla-sektora-finansowego/xdcw0>.
- [19] "Usługi biznesowe dla sektora finansowego", *Success story of Poland*. Report prepared by the Association of Business Service Leaders In Poland (ABSL), May 2013, p.8.
- [20] "Sektor nowoczesnych usług biznesowych w Polsce", Report prepared by the Association of Business Service Leaders In Poland (ABSL), 2014, p.30.
- [21] Target BPO, 2014. Available: www.targetbpo.pl.
- [22] BSS Business Support Solution, 2014. Available: www.bssce.com.
- [23] "Czestochowa. Nowa lokalizacja na mapie usług biznesowych", Report prepared by the Association of Business Service Leaders In Poland (ABSL), 2014, p.20.

Fourth dimension of spatial description in business processes

Cezary Stępniać

Abstract—The article is devoted to capture the time factor of descriptions of business processes. The issue in question is an extension of the problems of the use of descriptions of spatial tools (based on GIS technology - Geographic Information Systems) to describe the modeling and execution of business processes. This article assumes full development of business process models (not just the algorithm process, but also the semantic layer model, actors, documentation, necessary resources and performance indicators). Using standard tools (such as BPMN, UML), it can design business processes, and then deploy them in business entities. The solution proposed in the article assumes that it is possible to create a map of the organization, which will deploy the actors and resources available to them. The map is interactive and allows the registration of events taking place in the organization. All state changes resulting from the implementation of specific processes will be updated on a map of the organization. In this way you will be able to visualize the status of processes in any dimension of time.

Keywords—Time factor in Business Process Modeling, spatial description, GIS methodology Information System integration.

I. INTRODUCTION

MODELING business processes is one of the key elements in the implementation of process management in organizations [1], [2]. Developing algorithms of procedures should promote the increase in the efficiency of an organization. Assuming the Resources-Based Management, only modeling algorithms is not sufficient for efficient management [3]. In addition to the algorithms of process, it is important to define the actors involved, the required resources, created (required) documentation and performance indicators, both for the whole process, as well as individual operations. The aforementioned items should be treated comprehensively. This allows to create rules of organization and the support of corporate dictionary.

It seems that the first phase of the creation of process modeling tools has already passed. With the adaptation of the tools used to create computer systems and their adaptation to the needs of business process modeling, such as, among others, UML AD (Unified Modeling Language Activity Diagram) [4] and BPMN (Business Process Model and Notation) algorithms

building process is no longer a problem [5]. Moreover, modern tools enable dynamic matching of processes to the needs to occur and the analysis of conflicts between different version of the some process.

Increasingly challenging for those managing processes becomes overseeing the implementation of the current processes. The point is the examination of the effectiveness of individual processes and their fragments, used resources, the involvement of actors or analysis of existing conflicts, unused resource bottlenecks like [6].

One of the key factors to be considered in these studies is the time factor [7]. In retrospect, one can examine the number of processes, their quality and compliance with the criteria of evaluation, the results of individual units and others. In this study, the analysis of subjects can be many different types of data-object classes described in the information systems. This can cause difficulties in their presentation.

A wide range of different types of visualization methods was developed. They are available including Cocpit as manager in BI (Business Intelligence) tretien system as a part of the ERP (Enterprise Resources Planning) [8]. They can also enrich the business process modeling and visualization of the status of their implementation.

This article discusses the problem of visualizing the effects of business processes, taking into account the time factor. In this regard, the GIS methodology (methodology of Geographic Information System) to build maps of organization and maps of processes was used. The aim is to describe the processes and effects of examination of the ongoing processes. Its goal is to be the current description of the various types of organizational resources.

The main theme of the article is to show how the time factor can be used to describe the current situation of the organization. Description of the time factor will be based on data collected in the information systems organization and business process modeling tools.

II. RESOURCE-BASED APPROCH IN PROCESS MANAGEMENT

The success of modern enterprises depends on many factors. Considered to be the main factors are those associated with the skills of acquiring new customers and markets, or ability to enter into various types of economic activities. In other words, it is an appropriate setting for the reception of external factors and transforming them into new business opportunities.

The mentioned approach may be somewhat counter

C. Stępniać, Czestochowa University of technology, Faculty of management. Al. Armii Krajowej 36b, 42-201 Czestochowa, Poland (corresponding author to provide phone: +48 881-311-610; fax: +48 34 3 250 351; e-mail: cezary.stepniak@gmail.com).

Resource-Based approach [9], [10]. Resources of the organization should be optimized in terms of its production capacity. It is very difficult that the modern perception of the organization's resources has been substantially expanded. In addition to traditional material resources, raw materials and financial increasing, the attention is paid to the soft resources such as: knowledge resources, relational or logos [11].

Soft resources become an important success factor. The problem is that in some situations, those resources instead of being a factor in the success may become unnecessary ballast in organization. It is essential therefore the issue of productivity of the resources contrasted with the potential costs of obtaining them, if necessary.

Estimating the resources needed in the organization can be made in the course of business process modeling. The issue concerns the extended modeling comprising the following elements [12]:

- 1) The algorithm of the process,
- 2) Actors
- 3) Created documentation,
- 4) Identification of all the types of resources necessary for the performance of the process,
- 5) Performance of indicators of processes, fragments of processes and operations.

Modeling of the algorithm of a process creates outline organizational procedures. It specifies how to implement business processes. It should be assumed that the designer has developed a procedure for it in such a way that it is implemented as efficiently possible. The algorithms can be developed with the help of tools like UML AD and BPMN.

Algorithms can be applied to process as actors. For this purpose, it is possible to use, for example, DFC (Deployment FlowCharting). In this way, it is possible to identify the actors at any level of detail (specific employees, organizational units, branch offices, outsourcing, sub-contractors, etc.).

In developing process models, there should be semantic order of organization. Actors marked in the DFC should be adequate to the list of OC (Organizational Chart) or a dictionary of partners (cooperating entities) [13].

DFC type diagrams and OC have different functions and can be relatively difficult to integrate them into a common visualization purposes. Moreover, some actors can perform many operations within the same process. Visualization can be even more complicated when it will refer to the description of the current status of ongoing multiple processes.

The condition for the visualization of the processes is the current registration of the carried out operations. At the stage of the modeling process, documents or conditions of them are defined. They will perform the following operations. Registration takes place by means of appropriate information systems, eg. ERP or AOT (Automatic Offices Tools) integrated with ICT tools (especially the Internet solutions) [14]. During the implementation phase of the operation, the responsible entity will be required to note the operation within the framework of the process and to describe its current state.

As a result, the description of each operation will be a source of data [15], [16].

The fourth element of the description is a definite description of resources. It is a multi-task. The starting point is to develop a classification of resources within the organization. In the literature, there are many classification of the organization's resources. They are mainly theoretical in nature. That classification of resources in the organization is of a practical nature. It creates a certain type of index, and on the basis of dictionary resources. This makes it possible to describe those resources in information systems (mainly ERP, but also include a GIS or CAD - Computer Aided Design). This allows to define resource needs, make an inventory of existing, as well as their allocation within the accepted rules. With the classification of resources and leading their records, it is possible to assign them to specific operations modeled processes. These actions constitute the extension of the use Technological Cards in Production module of the ERP. This allows then to evidenciate the use of resources during the implementation process. From the point of view of resource organization, it is said that action allows to specify the consumption of certain types of resources or estimate the level of their use in relation to other types.

The fifth element is the performance indicators. With it, it is possible to define the potential requirements for the designed processes and their operation and to evaluate the implemented processes. For the development of performance indicators it is necessary to develop the rules and dictionaries. On this basis, a formula to calculate them was developed, enabling to enumerate them and make use of the various types of ratings organization's resources (including human resources). The results can also be used to modify process models and the introduction of various types of organizational changes.

Development enhanced process models will require additional organizational effort, but at the same time will sort dictionaries organization, the organizational arrangements and integrate different types of information systems around a common data. Despite the use of different modeling tools for individual items, it is important to maintain consistency of terminology on all models.

This makes possible semantic networking between different elements of the model and define existing relationships. Defining these relationships is a prerequisite for the construction of tools for comprehensive visualization of the modeled and implemented models [17].

III. TIME FACTOR IN BUSINESS PROCESSES

Time is an important factor in the success of any process [7]. It should be recognized both in the design of business processes and then in progress. A matter of time appears in the business processes in various aspects.

At the stage of modeling the time factor is seen mainly in three aspects.

The first aspect is the time of modeling. Process management is focused on the realization of its objectives,

which are usually associated with the desire to meet the needs of the customer. Often in the specification of the client's needs there is the time factor. It is the nature of limiting the duration of the project. This means that the designer of the process will not have unlimited time to develop a model of the process, and also with the process model should take into account the time factor specifying the conditions for implementation.

In the extended model, the time aspect of the process occurs mainly in two aspects: documentation and efficiency ratios. Each document is designed or used for describing of the processes that should be dated. Project's documents should be considered as a document dating the field, as well as any modification to it.

The time can be one of the determinants of the effectiveness of the proposed processes. Modeling can be determined the absolute time of implementation of the operation. Their transgression may result in penalties for actors implementing them. Penalties may be imposed automatically by the control functions of information systems supporting the realization of specific processes. You can also define a formula listing the duration of the whole process and the different activities depending on the specifications provided by the client.

Even more important is the importance of the time factor in the description of the processes. Time is important both for the current records of ongoing operations and processes, as well as for the purposes of reporting, analysis, control and planning.

As mentioned, each operation should be registered in the system. This allows to specify, for instance: status of implementation of the process, the actor as a contractor of the operation, duration of the operation, the resources involved, and to verify the compliance with the execution of the operation agreed performance indicators.

The data collected allows to make all sorts of analysis which could include issues like: the number of processes performed in a specific unit of time, the number of operations carried out by the various actors in a given period, the volume of waste or used resources, the effectiveness of ongoing operations and so on. The individual analysis can be made for any unit of time [18].

Business process models can also be used to plan future activities. It, inter alia, indicate the performers of the planned processes or to reserve the necessary resources for a specified period of time models.

IV. ASSUMPTIONS OF SPATIAL DESCRIPTION

Spatial description allows for visualization of different types of phenomena, processes, fragments of reality with the help of GIS methodology. Visualization can be done using a geographic area or heuristic (arbitrarily defined space with its own coordinate system and the logic of the allocation of objects).

Spatial visualization is a significant advantage. It allows the simultaneous presentation of many different types of class-object, specifying the weight of the objects presented within the accepted criteria and an indication of the relationship

between different types of objects [19].

For visualization is necessary to define:

- 1) Space - S ,
- 2) cartographic grid - G ,
- 3) classes of objects - $C(O)$
- 4) spatial attributes of individual objects - A_S
- 5) describing attributes of individual objects - A_D
- 6) symbolization rules,
- 7) data sources.

By using GIS methodology it is possible to define and integrate different types of space including the use of hypertext. In this way, the geographical spaces and the heuristic ones can complement each other, or visualize various aspects of the presented processes.

Cartographic grid determines the logic of the allocation of objects in space by creating a reference system and determining the dimensions of the presentation. With the adoption of universal cartographic grid resources it is possible to transfer data between different geographic GIS tools.

ClassObjects $C(O)$ defines layers on the maps [20]. These classes of objects can be defined in other classes such as information systems such as ERP and CRM (Customer Relationship Management). Different types of objects form separate layers. From the point of view of Resources-Based Approach, different types of resources may be distinct $C(O)$.

As a part of each ClassObjects, instances representing a single object are distinguished. Individual objects can participate in a number of events (for example, participate in the implementation of operations in different processes). Thus, their states can dynamically change with the participation in subsequent operations. That variability over time can cause changes in the values of selected attributes describing A_D . The mentioned variability means that in order to preserve the visualized objects it is necessary to keep the access to the DB (data bases) or DocB (DocumentBases) in which data is stored on the implementation of individual operations [21]

As mentioned, individual objects are determined by spatial attributes describing A_S and A_D . A_S spatial attributes are responsible for the locations of objects on a map [22]. They should be constant in the case of stationary objects and variables in the case of mobile. An example of variables A_S may be the GPS coordinates of the object.

A_D attributes describing the object determine the states of a given object according to the preset criteria. The values of these data can be a constant value (eg. Employee's date of birth), variable (number of completed courses) or be a function (eg. Sales value, which is calculated on the basis of invoices assigned to a given employee).

Spatial visualization can be developed for different purposes. For the same data, the set can generate many different maps. Therefore, in the construction of a particular map, it is important to indicate which layers are displayed, and according to which A_D will present individual objects. The use of interactive map allows to change layers (on some, off others), to change the symbolization of individual objects (in

the case of changes in the visual criterion $C(O)$, as well as responding to further entries made in the DB Or DocB under which individual objects are presented [23].

It should be noted that in the case of using interactive maps, the maps are visualized on-line with a change, making another entry in computer systems. The time factor in this case will be crucial for visualization of the presented phenomena. models.

In today's market, spatial information began to emerge as three separate but cooperating groups of actors [24]:

- 1) Providers of GIS technology
- 2) Spatial Data Providers
- 3) Spatial Analysts

This division means that GIS technology suppliers provide only a tool to help build a map and possibly basic thematic layers in the case of geographical space. To recover the maps, spatial data is necessary. Their collection and sharing deals with many subjects. Increasingly, they are available on-line via the Internet. In addition to the file formats strictly geographical spatial data, are increasingly turning to GIS. It can import data from publicly available software packages (eg. Office) or different types of systems (eg. ERP / BI, CRM, CAD) [25]. GIS technology expands the possibilities of entities engaged in the provision of tools for spatial analysis, aimed, inter alia, to model spatial phenomena of nature. While maintaining the spatial visualization of business processes, it is possible to build models of execution of business processes in specific time periods.

V. THE FOURTH DIMENSION

Application of the use of GIS technology has changed the rules of cartographic visualization. The existence of 2D maps were unchanged. It was only manually to change the data on a map or print its subsequent revised versions. To use the map it was necessary to carry it.

GIS Technology has allowed for the introduction of 3D by using transparent layers. In this way, it is possible to visualize in 3D reality providing that the relevant data is available (an example [26].

Theoretically, it is possible to create mathematically nD spaces. It is a question of defining the relevant A_s . However, the visualization can be carried out for technical reasons only in 3D. However, it is possible to switch axis dimensions, what change will also be presented to the map.

Generally, it is assumed that the fourth dimension is time. Modern GIS technology allows its use in spatial visualization. One can imagine an interactive map that shows the history of Europe, and actually change the borders of Europe in any defined period and speed of moving along the timeline (or also do back) [27]. The only requirement is to have the access to the relevant data.

The use of spatial visualization of descriptions of business processes is designed to provide a tool to facilitate the modeling process and handle their implementation. The proposed tool can also schedule during the execution of individual processes, for instance, by planning the use of

resources of the organization.

The present discussion is limited to the factor of time (the fourth dimension) in drawing tool used to of descriptions of business processes.

The starting point is to determine the importance of time in the modeling and implementation of business processes in the organization. Adequately to this you can choose the solution to use the issue of time.

If it is planned to use a specific time issues they should be considered in the three interrelated phases:

- 1) process modeling,
- 2) execution plans
- 3) recording and control of the effects of the processes.

In business process modeling time issues included in the project will focus on the documentation and efficiency ratios. Assuming the extended modeling of processes, the documents should be identified, the ones which will reflect the implementation of subsequent operations. From the technological point of view, each newly created document should be saved in the DB or DocB, to the appropriate system which will register and implement operations (usually this will be ERP). Each entry is also subjected to the dating and the user who made it should be identifiable [28].

When designing procedures it may be determined by the expected duration of individual operations and / or process as a whole. It should be assumed that the duration of the process or a fragment therefore can not be shorter than the sum of its parts operation times. It can also impose conditions limit on the time factor, imposing time constraints and specifying the conditions under which it makes it possible to refuse to accept the order (eg. due to too short lead time required).

Expected timing indicators for individual operations may implicitly define the time you book different types of resources assigned to specific operations.

Defining indicators of time means that for some systems will need to impose a verification procedure. The most common ERP systems, but also include in WMS (Warehouse Management Systems) or CRM can be built and control procedures counting the execution time of each operation based on the registration of relevant documents. These procedures will be recognized during the conversion of business process models for system procedures [29].

At the stage of planning processes the role of the time factor is increasing. The spatial visualization may also be used. When planning the execution time of specific processes, one of the essential elements of a resource management organization. Spatial visualization can be useful mainly when the forecast will be the implementation of multiple processes in parallel. Using the map of the organization with the presentation of the available resources and the use of intelligent technologies can be booked map specific resources to complete the process. Description the individual objects on the map can be combined with the technology hypertext so directly on the map allow applied the relevant records in databases describing the object. Dynamic Visualization may

indicate that the facilities will be available at an assumed within the process.

The application of these solutions will require the integration of GIS with ERP (or others), where it is recorded by a description of the required facilities, and business process modeling tools in which you saved the process model. GIS tools should have functions such as spatial simulation, visualization and intelligent periodic maps of hypertext.

The key benefits of the time factor can be achieved in the implementation phase and process control. Using the integration of GIS with DB (and DocB) of ERP, CRM, CAD-to-date, you can visualize the map with the registration documents describing projects following operations processes. Creating a process map, you can track the status of each of the processes. In addition, using GIS technology it is possible to overlay map of the processes. In this way, the involvement of individual organizational units (actors) in the functioning of the organization can be indicated. On the basis of the documents describing the implementation of the operation should be a description of the state of the resources used to implement the operation. Thanks to this it can be specified, inter alia, the state of resources and their consumption, availability and more. On this basis it is possible to indicate the usefulness of the resources or the level of consumption. With various types of data processing functions can be of different types of parameters to calculate, which could provide the attributes describing the AD for each object. United visualized resource can be presented on one definite point in time, on-line updated or animations can be carried their states at a given time, as well as the simulation of expected conditions in the future by using the assumed planning data.

Visualization of the spatial can also be used for control purposes. Assuming a specific period of time, any defects, deficiencies can be visualized that occurred in a given period of time. This allows to build an incentive system of the organization. Animations of specific processes, phenomena, or fragments of reality may indicate emerging negative or positive phenomenon and that outlines the trend. It seems that it is much easier to interpret than the map is a dynamic set of tables. The data can be collected in distributed DB systems [30]

With maps visualized it is possible also to find what is currently available, and whether they can be used (appropriate symbology may indicate that the resource is reserved within a specified period). Therefore, the application of the time factor significantly larger usefulness of the proposed tools.

The use of the proposed solutions require the use of appropriate technology. The starting point is the integration of different types of information systems (including GIS, ERP / BI, CAD / CAM, CRM, SCM). Appropriate interfaces must be able to communicate on-line with the wide area network (especially the Internet). This allows data stored in different systems can be imported into GIS and processed, and then visualized. GIS tools should provide appropriate analytical functions, taking into account the time factor allowing the

current animation and simulation predicted phenomena.

From the point of view of design and planning processes of resource use animation time (dynamic visualization of the level of use of specific types of resources) may be a factor in stimulating entrepreneurship. It can manifest itself new project business processes in order to develop the unused resources.

The use of the time factor is a relatively new development in order to visualize phenomena. So far, dynamic modeling was adjusted to simulate phenomena in geographic environments. In relation to business processes, the main problem is the need for multiple calculation procedures for different issues seems that in this case will be the future of the object-oriented technology DB followed for the development of GIS,

VI. CONCLUSION

The thematics taken into consideration in the article is a research section in the larger whole, whether it is the use of GIS technology in description of business processes. The main current research is thus aimed at the integration of different types of information systems, data formats and tools of ICT (Information Communication Technology). However, it is difficult in this time of a skip a factor in the development of the proposed tools.

Taking into account the time factor, will make it easier to visualize the phenomena or more dynamic business processes. It should also influence the perception of better presented maps.

Although theoretical considerations have character, their preparation was based on empirical research conducted with many interviews and research with representatives of different environments. Application basic problem stems from the difficulty of finding entities, which is, in conscious way, introducing management process undertaken by the extended business process modeling. Companies engaged in the development of GIS tools have just become the subject of the application of their products, not only for geographical purposes. Companies producing other systems (eg. ERP / BI) assume that their functions are analytical and planning level is adequate to the needs of the client and so far they are afraid to invest in the proposed technologies. Large traders so far not yet fully coped with the integration of all the applicable systems.

Therefore, the analyzes were carried out on the base of technological possibilities of these tools. In contrast, application options discussed should come to the end of the fourth dimension.

REFERENCES

- [1] J. Freund, B.R..Rucker, "Real-Life BPMN : Using BPMN 2.0 to Analyze, Improve, and Automate Processes in Your Company". Published by Createspace 2013.
- [2] M. Havey, "Essential Business Process Modeling" Ed. By O'Reilly 2005.
- [3] Jelonek D., Stepniak C., *Dynamic Business process Modelling in Organization*, In "People Knowledge and Modern Technologies in the Management of Contemporary Organizations. Theoretical and

- Practical Approaches*. Ed. By. Csaba Balint Illes and Felicjan Bylok. Goddolo 2013.
- [4] W. Dąbrowski, A. Stasiak, M. Wolski, „*Modelowanie systemów informatycznych w języku UML 2.1 w praktyce*”. Wydawnictwo Naukowe PWN SA. Warszawa 2007.
- [5] F. Schonthaler, G. Vossen, A. Oberweis, T. Karle, “*Business processes for Business Communities. Modeling Languages, Methods, Tools*”. Ed. by Springer.-Verlag Berlin-Heidelberg 2012.
- [6] Gerth C., “*Business Process Models. Change Management*”. Ed. By Springer.-Verlag Berlin-Heidelberg 2013
- [7] M. Laguna, J. Marklund, “*Business Process Modeling, Simulation and Design*”, Second Edition Ed. by Taylor & Francis Group LLC, 2013
- [8] Mendeleev table-types of visuasization. http://www.visual-literacy.org/periodic_table/periodic_table.html.
- [9] J.B. Barney, “*Resource-based theories of competitive advantage: a ten-year retrospective on the resource-based view*”. In *Journal of Management*, 27/2001, 643–50.
- [10] R. Krupski, *Orientacja zasobowa w badaniach empirycznych. Identyfikacja horyzontu planowania rynkowych i zasobowych wielkości strategicznych*, Walbrzyska Wyższa Szkoła Zarządzania i Przedsiębiorczości, Walbrzych 2011.
- [11] M. Mach-Król, *Narzędzia budowy systemu z temporalną bazą wiedzy wspomagającego twórczość organizacyjną*. W „*Informatyka Ekonomiczna, Business Informatics*” Nr 2(32) 2014. Wyd. UE Wrocław 2014 s. 179-187.
- [12] D. Jelonek, C. Stępnia, *IT Support for Resource - Based Approach in Enterprise Management. W: Contemporary Economies in the Face of New Challenges. Economic, Social and Legal Aspects*. Edited by Ryszard Borowiecki, Andrzej Jaki, Tomasz Rojek. Publishing House: Foundation of the Cracow University of Economic. Cracow 2013.
- [13] D. Howard, *The Basics of Deployment Flowcharting & Process Mapping. A User's Guide to DFC for Know-how Capture and Process Design*. Ed. Management New Style <http://www.flowmap.com/documents/booklets/dfc.pdf> (last view 12.03.2015).
- [14] R. Stair, G. Reynolds, “*Information Systems Essentials*”. Fifth Edition. Edited by Course Technology 2010.
- [15] D. Cohn, R. Hull, “*Business Artifacts: A Data-centric Approach to Modeling Business Operations and Processes*”, In “*Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*”. Volume 32, Number 3, September, 2009, pp. 3-9.
- [16] M. zur Muehlen, “*Volume versus Variance: Implications of Data-intensive Workflows*”. In “*Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*”. Volume 32, Number 3, September, 2009, pp. 42-47..
- [17] I. Pawełszek, *Technologie semantycznego Internetu w kreowaniu przedsiębiorczości nowej ery*. W „*Wiedza i technologie informacyjne w kreowaniu przedsiębiorczości*”. Monografia. Red. nauk. A. Nowicki, D. Jelonek. Sekcja Wydawnicza WZ Politechniki Częstochowskiej. Częstochowa 2013.
- [18] H.-L. Truong and S. Dustdar “*Integrating Data for Business Process Management*”. In “*Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*”. Volume 32, Number 3, September, 2009, pp. 48-53.
- [19] F. Harvey, „*A primer of GIS. Fundamental Geographic and Cartographic Concepts*”. The Guilford Press. New York London 2008.
- [20] P.A. Longley, M.F. Goodchild, D.J. Maguire, D.W. Rhind, “*GIS Teoria i praktyka*”. Wydawnictwo Naukowe PWN. Warszawa 2006.
- [21] Wierzchowski J.: *Big data – aspekt technologiczny i ekonomiczny vs. aspekt społeczny*. W *Ekonomiczno-społeczne i techniczne wartości w gospodarce opartej na wiedzy .t.2*, red. Jacek Buko, Zeszyty Naukowe nr 113 *Ekonomiczne Problemy Usług*, Uniwersytet Szczeciński, Szczecin 2014, s. 399-408,
- [22] C. Stępnia, „*Integracyjna rola przestrzeni informacyjnej w modelowaniu procesów biznesowych*”. W *Zeszytach Naukowych Uniwersytetu Szczecińskiego nr 702 - Ekonomiczne Problemy Usług nr 87*, Wyd. USz. Szczecin 2012, część I, s. 513-521
- [23] C. Stępnia, *Mapy interaktywne jako narzędzie wspierania procesów inwestycyjnych*. W „*Roczniki Kolegium Analiz Ekonomicznych*”. Oficyna Wyd. SGH. Warszawa 2015, to be Published
- [24] C. Stępnia, „*Kartograficzne rozszerzenie dynamicznego modelowania procesów biznesowych*”. *Zesz. Nauk. Uniwersytetu Szczecińskiego nr 808* Tytuł zeszytu: *Ekonomiczno-społeczne i techniczne wartości w gospodarce opartej na wiedzy*. T.1, Wyd. USz. Szczecin 2014, s. 441-448.
- [25] D. Jelonek, C. Stępnia, T. Turek, 2013, “*The Concept of Building Regional Business Spatial Community*”, In *ICETE 2013. 10th International Joint Conference on e-Business and Telecommunications. Proceedings*, 29–31 July 2013, Reykiavik, Iceland.
- [26] M. Pietruszka, M. Niedźwiedziński: *Third dimension of e-commerce W W „Informatyka Ekonomiczna, Business Informatics” Nr 2(32) 2014*. Wyd. UE Wrocław 2014, s. 198-212.
- [27] MapMaker <http://mapmaker.edukation.nationalgeographic.com/> (last view: 26.11.2014).
- [28] P. Polak, J. Wierzchowski, “*Rozwój metod modelowania procesów biznesowych dla potrzeb wytwarzania systemów informatycznych In Podejście procesowe w organizacjach*”, red. Stanisław Nowosielski, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 169, 2011, s.266-272
- [29] D. Jelonek, J. Nowakowska-Grunt, L. Ziara, *The Assessment of Construction Project Management Maturity Level in the Silesian Region in Poland*, *Advanced Materials Research*, 2014, Vol.1020, pp.796-802.
- [30] M. Tamer Ozsu, P. Valduriez, *Principles of Distributed Database Systems*. 3rd Edition, Springer Science+Business media. New York, Dordrecht, Heidelberg, London 2011.

Cezary Stępnia, PhD in Economics is an adjunct in the Faculty of Management at the Czestochowa University of Technology, Poland. His research interest include management of information systems, business and information processes The application of technology and geographic Information System of descriptions of management processes. He has near 100 publications as books, papers in conference proceedings and papers in journals.

Design and Implementation of the Korean Style Plug-In using the Wordpress

Jeongseok Ji, Jaesic Kim, Youngwan Kim, Sungjin Jung, Chaehyun Lee, Dongsu Kim, Yonggon Kim, Miyoung Bae, Yangwon Lim and Hankyu Lim(corresponding author)

Abstract— Following the advancement of IT technology, high quality open API market has emerged that are developed by outstanding developers. Especially, Wordpress is showing rapid growth in the open-source CMS market. Accordingly, the Korean CMS market is also experiencing change from the previous focus on XEengine into the Wordpress platform. Nevertheless, even though Wordpress market is growing in Korea and the number of homepages using Wordpress is rapidly increasing, it is still far behind the global market trend. The interpretation is that it is because UI in Wordpress does not fit the Korean sentiment. Hence, this paper developed Wordpress UI that is suitable for Korean sentiment by designing and implementing plug-in that supports application of Five-color and patterns useful for decorating blogs in traditional Korean style using Wordpress.

Keywords—wordpress, plug-in, korea style, CMS

I. INTRODUCTION

DOCUMENT is a Wordpress (<http://wordpress.org>) is one of the open-source CMS that has a characteristic of enabling free modification of website or menu composition even in case one is not familiar with HTML [1]. Wordpress is a writing tool that supports easy building and management of website and approximately 15% of the global websites are made using Wordpress. Moreover, since Wordpress is made in 'mobile response web design' that is compatible with mobile website, it has an advantage in that there is no need to build a separate webpage exclusively for mobile. Moreover, it is easily connected to social network service (SNS) such Twitter and Facebook. Hence, Wordpress is an interior tool that helps easy building and management of website and makes homepage easily seen on diverse mobile devices [2].

Homepage using Wordpress is composed of frame, theme, and plug-in. Currently, there are more than 7,500 plug-ins in official plug-in directory. In addition to this, a number of other supporting services are providing various themes and plug-ins. However, growth of theme or plug-in using Wordpress is stagnant in Korea.

II. RELATED RESEARCH

A. Wordpress Market Place

Figure 1 below presents the Wordpress usage and its market

This work was supported by a grant from 2014 Joint-industry-academic Research Fund of SMBA, Korea.

share as of January 2015 shown in W3Techs CMS trend. Wordpress turned out to occupy 23.3% of the total market based on the usage [1].

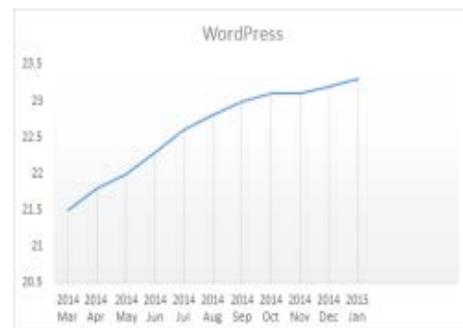


Fig. 1 Amount of CMS Usage

B. Wordpress Plug-In

Plug-in in Wordpress is a core component that enables users to freely change Wordpress as they wish without altering the core code of Wordpress. Plug-in is capable of realizing any imaginable function and most of the standard CMS works can also be processed using plug-in [3]. Plug-in in Wordpress is created using Wordpress API and the Wordpress can be limitlessly extended when plug-in is used. Moreover, in case of using plug-in, one can add function to Wordpress without altering the core code[4]. Hence, plug-in can be made either simple or complex according to the nature of tasks.

C. Five-Color

Five-color, also called as five colors of five directions, is traditional Korean color composed of five colors of yellow(黄), blue(青), white(白), red(赤), and black(黑). It is based on the concept of Yin and Yang that says the energy of Yin and Yang emerged to become heaven and earth and these two energies of Yin and Yang subsequently created the Five Elements of tree(木), fire(火), earth(土), gold(金), and water(水). Five Elements also involve five colors and directions. With the center and four directions as standard, yellow(黄) means center, blue(青) means East, white(白) means West, red(赤) means South, and black(黑) means North [5].

D. Plug-In Security

When creating plug-in, defending security against hacking or exploit is one of the most important stages. If security flaw of plug-in is exposed to malicious hacker, the whole website

composed of Wordpress can be disrupted [6].

For this, there is a security tool built in Wordpress that can enhance safety of plug-in. Nonces used in Wordpress means a number that can be used only once. In Wordpress, secret key is generated in order to determine whether diverse task requests (option storage, form template submit, Ajax request, etc.) are illegal and Nonces is used here. The secret key is generated prior to task requests, and the created key is transferred contained in the task request. The identity of the key is subsequently checked before running the script [3].

E. Wordpress Hook

In extending Wordpress function, hook is one of the important functions. Hook is a terminology used in Wordpress and, in general, it is called hooking. Hooking is a software engineering terminology which refers to command, method, technique or behavior that changes or interrupts in the middle in function call, message, or event that take place among software components in computer programs such as operation system or applications. In case of using hook, random function can run to change the Wordpress function or output in the middle of running Wordpress. Hook is the most important method that Wordpress contents and plug-in uses to exchange information and it can be called as 'PHP function call that has transmissible parameter' [7].

F. Action and Filter

There are two types of hook. An action hook is run in case of event within Wordpress. For example, action hook is the one that can be called when new post is registered. A filter hook is used when contents are stored in database or changed before being printed on screen. The filter hook involves contents of post or page. Hence, the contents can be changed after being transferred from the database or before being printed on the browser screen [7].

III. IMPLEMENTATION AND DESIGN

This paper designed and fabricated plug-in based on Wordpress platform. UI was composed using Javascript and CSS and color picker was fabricated using JQuery. As shown in Figure 2, an independent space was created in the same name for WordPress plug-ins to be stored in the WordPress installation folder "wp-content/plugins".

```

/wp-content/plugins/ks-fivecolor-plugin/
/wp-content/plugins/ks-fivecolor-plugin/ks-fivecolor-plugin.php
/wp-content/plugins/ks-fivecolor-plugin/inc
/wp-content/plugins/ks-fivecolor-plugin/css
/wp-content/plugins/ks-fivecolor-plugin/images
/wp-content/plugins/ks-fivecolor-plugin/js

```

Fig. 2 Plugin Folder of Wordpress

The "inc" folder collects other script files that will be used by retrieving from the main plug-in script file "ks-fivecolor-plugin.php". The "images" folder is created to

store relevant images such as icons and five-colored images. The "js" folder stores JavaScript files and the "css" folder stores style sheets. Once WordPress is executed, the "wp-content/plugins" folder is read in and all plug-ins installed inside the folder are identified, and then only activated plug-ins are executed. Not only the main system of WordPress, but also all plug-ins are executed within the same namespace. Therefore, as shown in Figure 3, the names of functions or variables were designed using a prefix.

```
ks_fivecolor_plugin_functionName();
```

Fig. 3 Function Name of Plugin

By using the prefix "ks_fivecolor", conflicts with other plug-ins can be avoided and PHP errors can be reduced. In addition, as shown in Figure 4, a plug-in header included in the main script file was configured.

```

<?php
/*
Plugin Name: Korean Style – Five Color Plugin
Plugin URI: http://thekoreansilk.com
Description: This plugin supports application of Five-color and
patterns useful for decorating blogs in traditional Korean style
Version: 0.0.9
Author: Korean Style Team, Andong National University
Author URI: http://multi.andong.ac.kr
License: GPL2
*/
?>

```

Fig. 4 Plugin Header

The plug-in header is used to obtain information related to plug-ins by parsing in WordPress, and plug-in information is printed out on the WordPress administrator's screen. The header is an essential item that should be included in plug-ins, and WordPress cannot recognize the plug-ins without the header. Plug-in files saved in this manner are compressed into a zip file, and can be used by installing them as plug-ins and activating them.

Figure 5 below shows pattern setting screen using plug-in that uses Wordpress developed in this paper. When setting the plug-in's pattern, background color and foreground color can be determined and the registered pattern image can be chosen to be applied. For development of Korean style plug-in, a function that enables setting of Five-color was added differently from previous color pickers and patterns were also designed such that diverse Korean style patterns can be included.

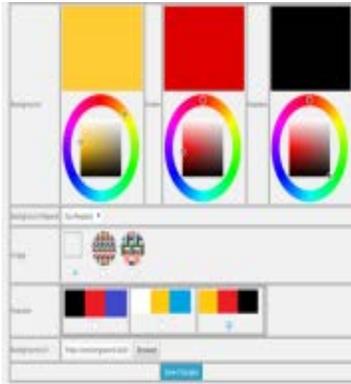


Fig. 5 Color-Picker Prototype

Figure 6 below shows screen where Five-color was applied. Korean style color picker plug-in that supports Five-color was designed such that background, pattern image, Five-color, etc. can be changed.



Fig. 6 Choosing the Background Color

It was implemented such that background can be repetitively set after choosing the background color or pattern image, by applying Background-Repeat. Moreover, it was composed so that it can be applied to header and footer of Wordpress.

Figure 7 and 8 below show a screen where color picker that supports Five-color is applied in initial layout of Wordpress



Fig. 7 Applied Pattern



Fig. 8 Applied Five-color

In this paper, plug-in was designed by applying Korean style Five-color, whose provision enables application of Korean style patterns, which was impossible in previous color pickers. In previous plug-ins, Korean style pattern could not be provided or Five-color could not be selected. The plug-in implemented in this paper enables controlling the blog's theme using diverse Korean patterns. Not only the background color, headers color and footer color can be freely chosen, but also the colors can be chosen following the Korean traditional color of Five-color in a differentiated manner.

IV. CONCLUSION

In recent CMS market, Wordpress became the number one domestic CMS tool, surpassing the express engine (XE). Following such trend in domestic market, the number of domestic homepages that use Wordpress is also rapidly increasing. However, there was no case in Wordpress market where theme or color that has exclusively Korean color or Korean sentiment was applied.

In this situation, this paper designed and implemented plug-in that enables making blog in Han(한)guk style by using Wordpress whose usage is ever increasing. Previous plug-ins and themes were mostly fabricated according to the international style. Here, Han(한)-style plug-in was fabricated, being differentiated from the previously made theme changing plug-ins. Considering that literature and information about Wordpress in Korea is still insufficient, there can be some difficulties when designing and implementing Wordpress in Korean style plug-in and theme.

Wordpress based on open-source enables many developers to easily develop additional extended programs, supporting limitless extension. Based on the plug-in that uses Korean color developed in this paper, it is expected that more plug-ins that fit Korean sentiment will be developed in the future. Moreover, in addition to the Korean market, its usability as a global plug-in is also expected to grow considering increasing amount of global attention to Korea these days.

REFERENCES

- [1] <http://w3techs.com/>
- [2] Griffin, Jonathan. "WordPress 4.0 New Features." (2014).
- [3] Brad Williams, Ozh Richard, Justin Tadlock, Professional WordPress Plugin Development, Wrox, 2011.
- [4] Hedengren, Thord Daniel. Smashing WordPress: Beyond the Blog. Vol. 32. John Wiley & Sons, 2012.
- [5] Byungwoo Kwak, A Study about Color Consciousness about Korean Traditional Five Colors in Art Therapy History and its Therapeutic Applicability, Youngnam University, 2011.
- [6] Canavan, Tom. CMS Security Handbook: The Comprehensive Guide for WordPress, Joomla, Drupal, and Plone. John Wiley and Sons, 2011.
- [7] Brazell, Aaron. WordPress Bible. Vol. 726. John Wiley and Sons, 2011.

Jeongseok Ji, Jaesic Kim, Youngwan Kim and Sungjin Jung are students of Multimedia Engineering Department of Andong National University, Korea.

Chaehyun Lee, Dongsu Kim and Yonggoon Kim are employees of Webonomics company, Korea.

Miyoung Bae is a PH.D. student of Multimedia Engineering Department of Andong National University, Korea.

Yangwon Lim is a full time lecturer of Multimedia Engineering Department of Andong National University, Korea.

Hankyu Lim received the B.S. degree in Electronics Engineering from the Kyungbook National University in 1981. He received the M.S. degree in Computer Engineering from the Yonsei University in 1984. He received the PH. D. degree in Computer Engineering from the Sung Kyun Kwan University in 1997. He is a professor of Andong National University.

A Comparison of Open-Source CMS - Focused on the CMS Market Place in Korea -

Yangwon Lim, Youseck Yang, Hyeonpyo Hong, Geunwoo Ahn, Jeongwoo Lee, Eunju Park,
Jihyeon Hwang, Yonggoon Kim and Hankyu Lim(corresponding author)

Abstract—Contents management system is a system used for managing pictures, voices, electronic documents, and other similar computer files. Following advancement of the Internet, diverse CMS for building webpage have been developed. Among the CMS used for building homepage, this paper compared the market shares in Korean and international CMS markets and conducted analysis on the performance, technology, and usability of Korean and international open-source CMS. Although open-source CMS market is expected to grow further, development that considers commonly used way in Korea will be necessary for the growth of open-source CMS in Korea.

Keywords—OpenSource CMS, CMS, Wordpress, Plug-in Software

I. INTRODUCTION

FOLLOWING the spread of computer and activation of the Internet in the 20th century, numerous homepages were born, provoking development of software that helps developers easily build homepage.

The number of people other than professional developers that open homepages has also increased and the market size of the Internet shopping mall recorded 19 trillion KRW in 2010 [1]. CMS for building homepage makes it possible to manage contents in any location as it supports web-based management and the utilization of open-source CMS is especially increasing these days [2].

In order to examine the current situation of open-source CMS that is becoming widely used, this paper compared characteristics as well as pros and cons of open-source CMS between domestic and international cases and analyzed their market share and usability.

II. RELATED RESEARCH

XE (XpressEngine) is a CMS that was once widely used in Korea, but the CMS that has largest market share worldwide is Wordpress.

A. XpressEngine

Among homepage building tools that are provided in CMS type to developers in Korea, XE (XpressEngine) is best known. Although XE was originally a web program that generates and manages BBS (bulletin board system) made in PHP language, it

This work was supported by a grant from 2014 Joint-industry-academic Research Fund of SMBA, Korea.

currently provides diverse homepage templates other than bulletin board and enables simple development of homepage using tool kit. That is, it supports easy fabrication of homepage without directly implementing diverse functions through web programming.

B. Wordpress

Wordpress, which is the largest open-source CMS in the world, was founded by Matt Mullenweg in 2003. It is one of the installation-type blogs that is intuitively constituted so that users can easily understand it after using for a while. It has an advantage of excellent flexibility in building homepage and utilization of relatively more themes and open-sources compared to other tool. As of July 2014, there are as many as 2,570 themes and 31,435 plug-ins registered in Wordpress official homepage [3]. However, it has disadvantages of limited design option, vulnerable security such as weak management of large-scale contents, etc[4].

C. Plug-in

In open-source CMS, diverse plug-ins exist for each tool. Plug-in is a core component that enables users to change shapes or functions as they wish without altering the core code of homepage building tool. Most CMS supports plug-in, but manual differs in each plug-in and a series of installation, activation, setting and using are required in most cases[5].

D. Naver Syndication

Naver syndication is an API service developed by Naver Inc. in Korea in April 2010. This API service is an API that defines synchronization rule between the website that contains contents and searching service that searches for contents. By addressing the disadvantage of previous method used in searching contents collection, burden on independent sites can be decreased while enhancing the quality of searching service. Moreover, in case of using Naver syndication, searching accuracy increases as the independent site's contents are updated in real time and formalized web documents are collected [6].

III. MARKET SHARE ATTRACTION OF OPEN-SOURCE CMS

Figure 1 below shows CMS market share in Korea. As of October 2013, Wordpress has the largest CMS market share in Korea with its proportion 41.0%. Considering that CMS market share of XE (eXpress Engine) exceeded 60% in 2011, it is clear that the market is very rapidly changing. This indicates the fact that the Korean market is transforming from a local market to a

global market.

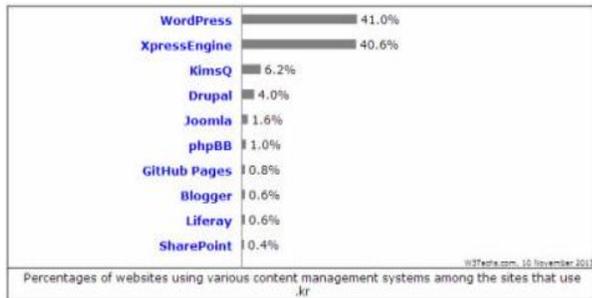


Fig.1 CMS Market Share in Korea

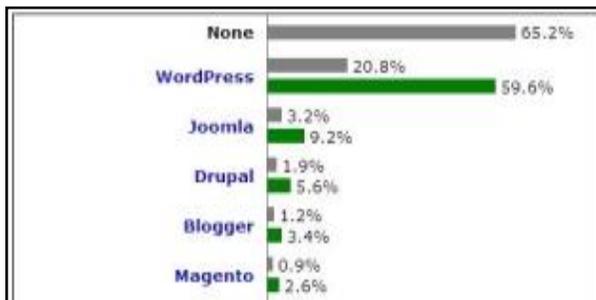


Fig.2 CMS Market Share

Looking at the international CMS market share, Wordpress has the number one CMS market share with 59.6%, which is far from Joomla that has second-largest market share (9.2%). Moreover, as is shown in grey bar graph in Figure 2 that indicates absolute usage percentage, one out of five global websites are made using Wordpress [3].

IV. A COMPARISON OF OPEN-SOURCE CMS

A. Comparative Analysis of the Open-Source CMS

Table 1 below shows the comparison results between domestically and internationally used CMS that were chosen according to their market share. Following the exponential growth of Smartphone since 2010, the trend has changed such that CMS can support response web. As of now, every CMS

turned out to be able to support the response web and search engine optimization program of SEO[7]. Except for the domestic CMS of XE and Gnuboard, all the CMS tools are capable of building shopping malls by installing separate plug-in. Especially, open-source based Magento is independently supporting building of shopping mall site and it supports all languages and currencies worldwide as well as major global payment methods. As of now, it is used in 1.5 million shopping malls, which constitutes 35% of the current online shopping mall solution market [8,9,10].

As for the multilingual support(MNLS:Multi-National Language Supplyment), Wordpress, Joomla, Drupal, etc. supported all the languages worldwide, while other tools had multilingual supports with limited languages. Wordpress turned out to have considerably larger number of plug-in supports compared to other tools and Magento had fewest. Naver syndication is supported by the Korean CMS KimsQ, Gnuboard, and XE. As Korean Naver is not well known in international CMS market, it was not supported. However, it can be used in Wordpress and Drupal by making use of separate plug-in and module. As for the program language and database, all the seven tools were using PHP and MySQL, while XE and Drupal turned out to be capable of using OracleDB, MS-SQL, etc., in addition to MySQL.

B. Comparative Analysis of the Open-Source CMS

Plan-on planner Kim Beom-soo once mentioned that “when strong member management function is required, the domestic CMS XE based on bulletin board is more proper than Wordpress” [3]. Internationally developed CMS is quite exotic in its structure or services. Although this can become an advantage in creating global service, it can put limits in case of building service for Koreans on the other hand.

Korean communities are mostly based on bulletin board. Large communities such as ‘DCinside’ and ‘Today’s humor’ are also based on bulletin board. When managing members in community, the members are classified or divided into sub groups according to their activities on bulletin boards, which are converted to scores. However, in case of using overseas CMS, although bulletin board can be made, there is no function that delicately manages members as in Korean community websites.

Table. 1 Comparative Analysis of the Open-Source CMS

	CMS	MNLS	SEO	Market share (%)		Shopping mall	Plug-In	Security	Dev Language	Database	Responsible Web
				Korea	World wide						
World wide	wordpress	○	○	41.0	59.6	○	○	X	PHP	MySQL	○
	Joomla	○	○	1.6	9.2	○	○	X	PHP	MySQL	○
	Magento	△	○	0.1	2.6	○	△	X	PHP	MySQL	○
	Drupal	○	○	4.0	5.6	○	○	○	PHP	MySQL, SQLite, Oracle	○
Korea	XE	△	○	40.6	0.1	△	○	○	PHP	MySQL	○
	KimsQ	△	○	6.2	0.1	○	○	X	PHP	MySQL	○
	GNUboard	△	○	0.4	0.1	△	○	X	PHP	MySQL	○

Another reason is font and graphical factors. Koreans actually put much emphasis on 'cute' fonts of homepage and they tend to concentrate on its graphical factors. Comparing the products displayed on three mostly used international shopping malls and top three Korean shopping malls, the difference is apparent. While international shopping malls have more text factors than graphics, Korean shopping malls have fewer texts. In the latter, the texts are either altered in images or more focus is put on the products' appearance and visual effects. Figure 3 below shows product description pages on one of the famous Korean shopping malls and on eBay. In case of many text factors, broken text phenomenon does not occur even if the page is zoomed in. However, the focus on the product description can be dispersed. On the other hand, in case of many graphical factors, jaggging can occur in graphic-type text when the page is zoomed, but focus can be concentrated on important product description.

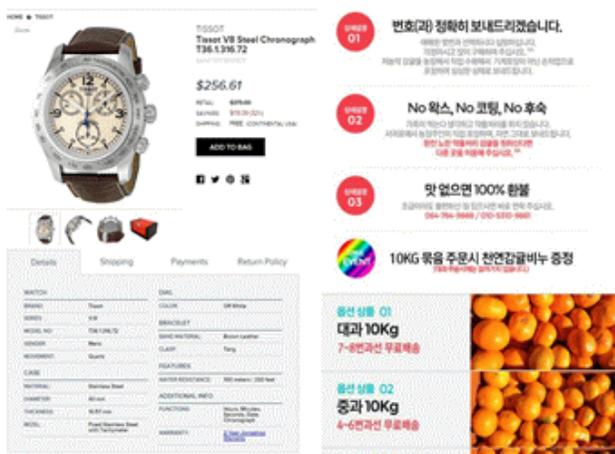


Fig.3 Comparison of ebay.com and auction.co.kr

Therefore, Korea's Web service market requires membership management functions to manage communities, and the open source CMS is required to develop and support plug-ins to compensate for those functions. However, product information on shopping sites operated in Korea tend to process even text information as graphics. Therefore, if the open source CMS, which basically supports the recent responsive Web, is implemented, it might not be able to respond to various mobile devices. For this reason, contents should be produced and improved by dividing them into text information and graphical elements.

V. CONCLUSION

Use of Wordpress in Korea is continuously increasing with an establishment of Wordpress User Forum in 2012. In current situation in Korea where the Smartphone penetration rate surpassed PC penetration rate, development of CMS response web has become more important. More progresses are expected of Wordpress that supports 'mobile response web design' and enables easy extension by many developers as it is based on open source.

This paper conducted comparison analysis on the characteristic and performance of Wordpress, by comparing the domestic and international cases, for the purpose of investigating the current situation of Wordpress. It turned out that CMS used in Korea possesses Korean unique characteristics. For introduction of more international CMS to Korea, it will be necessary to develop bulletin board and member management contents in a way that is commonly used in Korea, along with font or graphical factors development. To this end, this study plans to research Korean-style plug-in development technologies and guidelines on the production of contents that will be placed into the CMS.

REFERENCES

- [1] *Market Analysis(2004-2010)*, KISDI(Korea Information Society Development Institute), 2011.
- [2] Shreves, Ric. "Open Source CMS market share." white paper, Water&Stone, Summer, 2008.
- [3] Sangwook Ahn, *WordPress, Five charming three kinds of limits*, bloter.net, 2014
- [4] Walden, James, et al. "Security of open source web applications." Proceedings of the 2009 3rd international Symposium on Empirical Software Engineering and Measurement. IEEE Computer Society, 2009.
- [5] Williams, Brad, Ozh Richard, and Justin Tadlock. Professional WordPress Plugin Development. Wrox Press Ltd., 2011.
- [6] *Naver Developer Center, Syndication API Service Open*, 2010 [Online]. Available: <http://www.ddaily.co.kr/news/article.html?no=61445>
- [7] W3Techs.com
- [8] *Smartphones appeared four years ... '40 years'*, PC penetration pass, 2015, [Online]. Available: <http://10korea.com/smartphones-appeared-four-years-40-years-pc-penetration-pass/>
- [9] *Introducing Magento Open Source online shop solution.*, 2013, [Online]. Available: <http://runean.com/magento/>
- [10] ranker.com

Yangwon Lim is a full time lecturer of Multimedia Engineering Department of Andong National University, Korea.

Youseck Yang, Hyeonpyo Hong, Geunwoo Ahn and Jeongwoo Lee are students of Multimedia Engineering Department of Andong National University, Korea.

Jihyeon Hwang and Yonggoon Kim are employees of Webonomics company, Korea.

Eunju Park is a PH.D. student of Multimedia Engineering Department of Andong National University, Korea.

Hankyu Lim received the B.S. degree in Electronics Engineering from the Kyungbook National University in 1981. He received the M.S. degree in Computer Engineering from the Yonsei University in 1984. He received the PH. D. degree in Computer Engineering from the Sung Kyun Kwan University in 1997. He is a professor of Andong National University.

Support for reports and forms printing in wxWidgets GUI toolkit

Michal Bližňák, Tomáš Dulík and Roman Jašek

Abstracts—Despite its maturity and wide range of built-in features, the well-known cross-platform GUI toolkit called wxWidgets still lacks support for easy printing of reports and forms even in the latest version 3.0.2 released in October 2014. Although there are several technologies provided by the library which could be used for this purpose (like HTML-based forms) none of them are targeted directly to the printing process with all needed and expected functionality. This paper introduces new library add-on called wxReportDocument which fills this gap and allows users to easily create rich reports and forms able to preview and/or print run-time application data.

Keywords—wxWidgets, cross-platform, reports, forms, printing, data, binding, C/C++

I. INTRODUCTION

WELL-KNOWN cross-platform GUI toolkit called wxWidgets offers wide range of features covering nearly all functionality of target operation systems including GUI definition, processes and threads control, file system access, sockets, streams and many other features [2][5]. Unfortunately, dedicated support for printing of forms or reports is still missing even in the latest version of the library (at the moment the version number is 3.0.2 released in October 2014). Although there exists general support for printing allowing users to "manually" draw onto the printer's canvas (including print preview functionality) and possibility to print HTML-based content [3], there is no easy way how to defined full-featured reports and forms able to publish run-time application data stored in scalar variables or arrays.

This article introduces new wxWidgets add-on library called *wxReportDocument* created at Tomas Bata University aimed to fill this gap. The library offers a functionality needed for creation of print reports and forms. User-defined forms can be printed by using the library's internal printing back-end based on the wxWidget's printing framework or it can be saved to an output XML file (via any available output stream) for further usage. Also, it allows users to define unlimited number of printed pages consisting of various, highly customizable page items as shown later in this article.

The work was performed with the financial support of the research project NPU I No. MSMT-7778/2014 by the Ministry of Education of the Czech Republic and also by the European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089.

Michal Bližňák is with the Tomas Bata University, Faculty of Applied Informatics, Department of Informatics and Artificial Intelligence, Nad Stranemi 4511, 76005 Zlin, Czech Republic (corresponding author to provide phone: 00420-576035187; e-mail: bliznak@fai.utb.cz).

Tomáš Dulík and Roman Jašek are with the Tomas Bata University, Faculty of Applied Informatics, Department of Informatics and Artificial Intelligence, Nad Stranemi 4511, 76005 Zlin, Czech Republic (e-mails: dulik@fai.utb.cz and jasek@fai.utb.cz).

II. WXREPORTDOCUMENT LIBRARY'S INTERNALS AND STRUCTURE

The *wxReportDocument* library is created upon C++ implementation of wxWidgets GUI toolkit as its feature add-on. It uses built-in classes for graphics output, printing and previewing support and creates upper-level interface for definition of forms and reports. It can be used with both main wxWidgets 2.8 and 3.0 branches and can be easily obtained from wxCode add-ons repository [1].

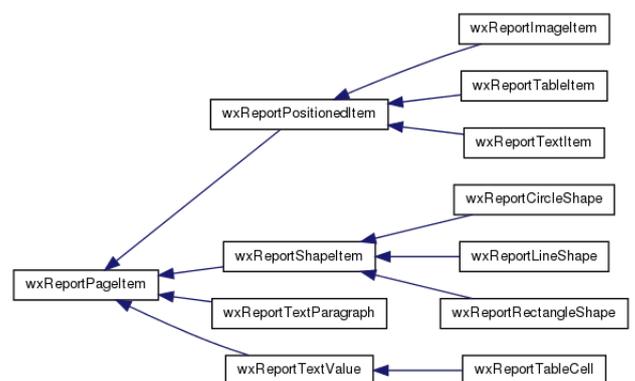


Fig. 1 inheritance diagram of page item classes

The library consists of the following main components:

Report document is encapsulated by *wxReportDocument* class which can be regarded as a main library class responsible for management of defined document pages. It also defines API functions for serialization, printing and previewing of the document. Typically, the report document class instance should be created and initialized at the application start-up and should be available during all application's life-time.

Report page is a basic printing entity managed by the *report document* encapsulated by *wxReportPage* class. Each the document can contain one or more report pages. Moreover, one report page defined by the user can be divided into several printed pages automatically when needed (e.g. when a table placed onto specific report page exceeds its vertical dimension) so the overall number of printed pages can be higher then number of user-defined pages.

Report page item is a single graphic object defined by the user and placed onto the *report page* which is encapsulated by *wxReportPageItem* class or another inherited ones. It can be

a static graphics object like bitmap image and vector shape (rectangle, circle or line), static text object defined in the source code, dynamic text object displaying textual representation of values created at the run-time, statically or dynamically defined tables or another user-defined item with custom drawing.

The Figure 1 show inheritance diagram of all available report items supported by the RP library.

Report style is a non visual object defining attributes of *report page* or *report page item*. It is encapsulated by `wxReportStyle` class whose objects can be assigned to report page items or directly to document page. It allows definition of foreground and background color, font, border style and other properties.

The Figure 2 show inheritance diagram of all available report styles supported by the RP library.

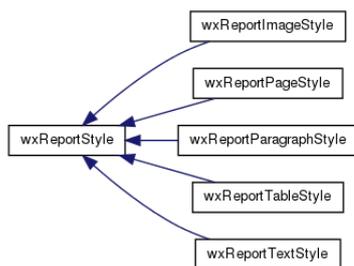


Fig. 2 inheritance diagram of report style classes

In addition to the mentioned components and classes the `wxReportDocument` library includes also other auxiliary classes encapsulating needed functionality. In the most cases, these classes are not intended to be used directly by the end-user so we will not discuss them here in details. For more information about them please refer to the library's project documentation available from the distribution package.

III. DEFINITION OF THE REPORTS AND THE FORMS

The following chapters deal with specific tasks needed for successful document creation.

The first step needed for the proper report document creation is definition of a report document object and setting its properties. As mentioned above, the report document object should be persistent for whole the application's life-time due to possibility to ask for the refresh/drawing of the document anytime later. Of course, this functionality is highly dependent on specific usage scenarios so it is completely up to the user how the report document object will be treated.

Now let us focus to the report page creation process. It is supposed that the reader is familiar with basic aspects of programming with C++ language and with usage of `wxWidgets` library as well. Also assume that global instance of `wxReportDocument` class named `m_Report` already exists like shown in Listing 1.

Listing 1: Main document report object

```

1 #include <wx/report/reportdocument.h>
2
3 wxReportDocument m_Report;
  
```

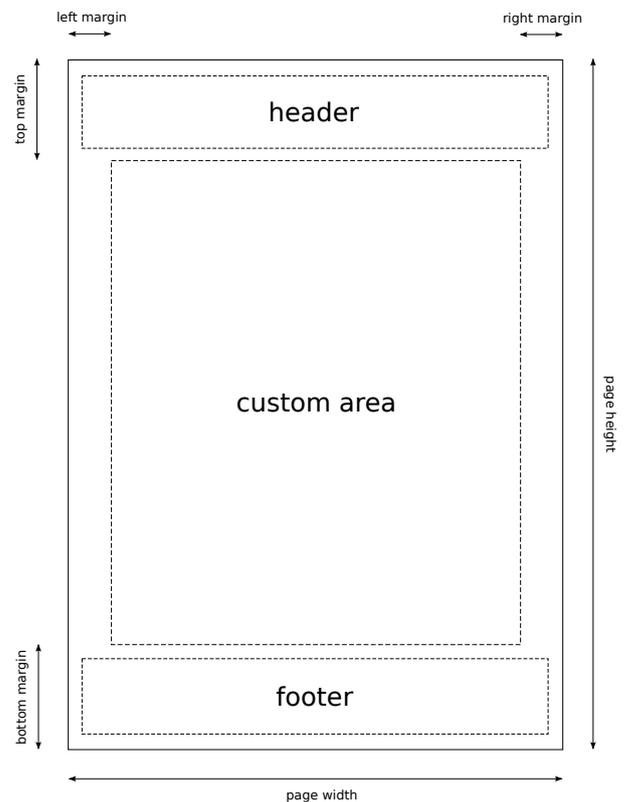


Fig. 3 basic page layout

A. Basic page layout

After the creation of the document object basic properties including the page size and margins can be set. Also notice that one document page is created at the document object initialization by default so it is not needed to do that explicitly.

Now, let us to discuss basic page layout as shown in Figure 3.

The report page consists of three main parts: *header*, *custom page area* and *footer*. The page width, height and margins surrounding the custom area can be set by using `wxReportPageStyle` class as can be seen from Listing 2. Remember that `wxReportDocument` library uses millimeters [mm] as its native dimension units type for all sizes and positioning.

Now, let us to define dimensions of the page and its margins.

Listing 2: Page layout definition

```

1 wxReportPageStyle pgs;
2 pgs.SetWidth( 210 );
3 pgs.SetHeight( 297 );
4 pgs.SetMargins( 10, 10, 30, 30 );
5 pgs.SetBorder( wxRP_ALLBORDER );
6 pgs.SetBackgroundColor( wxColour( 220, 220, 220 ) );
7
8 m_Report.SetPageStyle( pgs );
  
```

Defined page can be previewed by using special API function called `ShowPrintPreview()` defined in `wxReportDocument` class like shown in Listing 3.

Listing 3: Show report in preview window

```

1 m_Report.ShowPrintPreview( this, wxSize( 640, 850 ) );
  
```

As a response the application shows print preview window displaying defined pages and allowing user to print the content

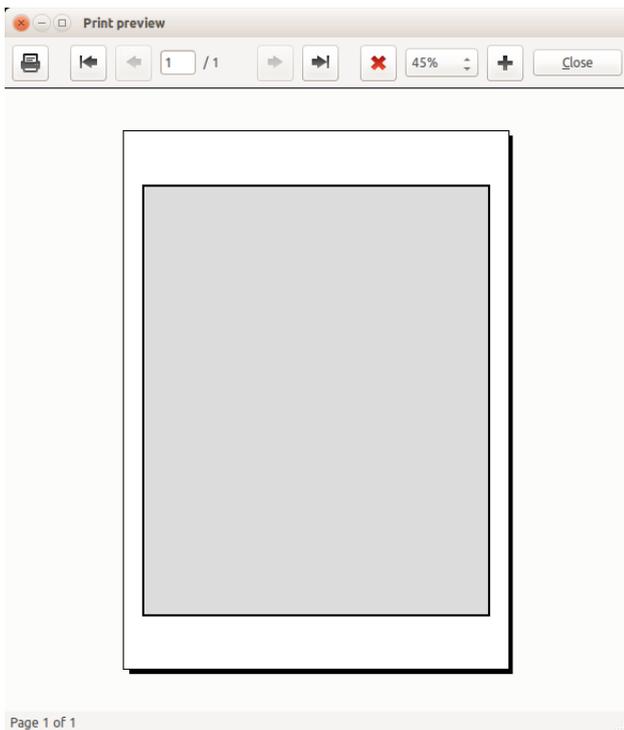


Fig. 4 print preview window

by using available printers as can be seen from Figure 4. Also, the document can be printed immediately without showing the preview window by using `wxReportDocument::Print()` function.

Headers and Footers In contrast to page items placed into custom page area, items managed by the header or the footer are not restricted by specified page dimensions and margins. It means that given positions of header/footer items are absolute while standard page items coordinates are relative to origin of the custom page area. In general, the content of the header, the footer and the custom page area can be the same. The difference is that a content of the header and the footer is repeated on all document pages while content of the custom page area can be unique. Also, there exists dedicated function for placing the page items into the header/footer declared in `wxReportDocument` class.

Listing 4: API for definition of the header/footer content

```
1 void wxReportDocument::AddItemToHeader(const
  wxReportTextItem& textItem);
2 void wxReportDocument::AddItemToFooter(const
  wxReportTextItem& textItem);
```

The following code shown in Listing 5 illustrates how to add some content into the page header.

Listing 5: Definition of the header content

```
1 wxReportTextItem hi;
2
3 hi.SetSize( 200, 20 );
4 hi.SetTextAlign( wxRP_CENTERALIGN );
5 hi.SetPosition( wxRP_CENTER, 10 );
6 hi.AddText( "Lorem Ipsum ..." );
7
8 m_Report.AddItemToHeader( hi );
```

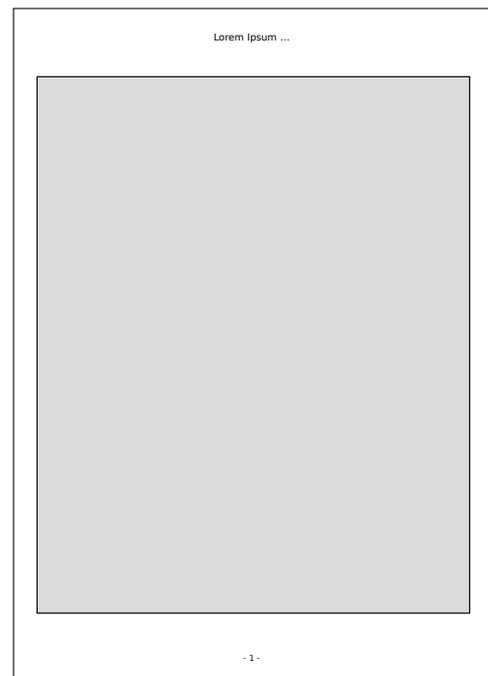


Fig. 5 numbered page with header

Page numbers The main class encapsulating printed document includes also API function called `InsertPageNumbering()` suitable for specification of page numbers. Of course, it is also possible to do that by using user-defined text items placed into the headers/footers but the dedicated function does lot of needed work itself.

The `InsertPageNumbering` function allows users to specify:

- format of page number in `printf`-like way,
- text style used with the page number,
- X- and Y- coordinates of the page number (in absolute values),
- target container (i.e. `wxRP_HEADER`, `wxRP_FOOTER` or `wxRP_BODY` respectively) and
- starting page.

Let us to define page number placed into the footer and show the result in Figure 5.

Listing 6: Definition of the page number

```
1 wxReportTextStyle ns( "ts_numbering", wxFont(10,
  wxFONTFAMILY_SWISS, wxFONTSTYLE_NORMAL,
  wxFONTWEIGHT_NORMAL) );
2 m_Report.InsertPageNumbering( "-%d-", ns, wxRP_CENTER,
  285 );
```

IV. REPORT STYLES

In the most cases, the library works in some sort of state-tracking mode which means that the defined styles are applied on all items added to the pages later after the style's definition and assignment. There are several classes encapsulating style

objects as shown in Figure 2 which can be used for specific needs. For example, `wxReportStyle` is base common style object defining mainly page/items borders and background color, `wxReportPageStyle` class can be used for definition of overall document look, `wxReportTextStyle` class can be used for styling of page text items, `wxReportParagraphStyle` class can be used for styling of text paragraphs, etc.

The style objects can be assigned to relevant page items or to the document page itself and can be re-used freely. Note that some page items can use several style object at the same time. Let us explain the styling process on text item encapsulated by `wxReportTextItem` class.

Text item can contain several words or lines with user-defined line endings and can be also divided into one or more paragraphs. The text item uses two different style objects: the *text style* and the *paragraph style* which are inherited from base *report style*. Except the common properties such are border style, border and background color the text style allows users to defined used font and foreground color while the paragraph style allows definition of text alignment, indentation, line height and paragraph spacing. Both styles can be assigned to the text item concurrently and are valid until replaced by another style object.

Now, let us to examine how the style objects assigned to the text item can affect its look.

At the first a text item consisting of three paragraphs with standard *Lorem Ipsum* content will be defined and inserted into the document page as shown in Listing 7

Listing 7: Definition of text item

```

1 float width = pgs.GetSize().x - pgs.GetLeftMargin() - pgs.
  GetRightMargin() - 10;
2
3 wxReportTextItem ti;
4 ti.SetPosition( wxRP_CENTER, 5 );
5 ti.SetSize( width, 0 );
6
7 /* assume existing lorem_ipsum_par_1, lorem_ipsum_par_2
  and lorem_ipsum_par_3 string variables containing
  printed text */
8 ti.AddText( lorem_ipsum_par_1 );
9
10 ti.AddNewParagraph();
11 ti.AddText( lorem_ipsum_par_2 );
12
13 ti.AddNewParagraph();
14 ti.AddText( lorem_ipsum_par_3 );
15
16 m_Report.AddItem( ti );

```

The text item's dimensions are calculated and specified at lines 1 and 5. Notice that just the width must be set by the user here - the height will be (in this case) calculated by the library itself. After that, the text item is center on the page within previously defined margins. The vertical position of the text is 5 millimeters under the top page margin. The content is divided into three paragraphs (see lines 10 and 13) and inserted by dedicated function at lines 8, 11 and 14. Finally, the newly defined text item is added to the page at line 16.

Now let us to play with the styles little bit. In the following sample, three style objects are used to style paragraphs, text content and the text item itself.

Listing 8: Definition of styled text item

```

1 float width = pgs.GetSize().x - pgs.GetLeftMargin() - pgs.
  GetRightMargin() - 10;
2
3 // define common report style

```

```

4 wxReportStyle rs;
5 rs.SetBorder( wxRP_ALLBORDER, wxColour(245, 245, 245) );
6 rs.SetBackgroundColor( rs.GetBorderColor() );
7
8 // define text style
9 wxReportTextStyle ts;
10 ts.SetFont( wxFont(12, wxFONTFAMILY_ROMAN,
    wxFONTSTYLE_ITALIC, wxFONTWEIGHT_NORMAL) );
11
12 // define paragraph style
13 wxReportParagraphStyle ps;
14 ps.SetTextAlign( wxRP_CENTERALIGN );
15 ps.SetParagraphsSpace( 5 );
16 ps.SetBorder( wxRP_ALLBORDER, *wxBLACK, 0.2 );
17
18 // define text item
19 wxReportTextItem ti;
20 ti.SetPosition( wxRP_CENTER, 5 );
21 ti.SetSize( width, 0 );
22
23 // set all styles
24 ti.SetStyle( rs );
25 ti.SetActiveTextStyle( ts );
26 ti.SetActiveParagraphStyle( ps );
27
28 ti.AddText( lorem_ipsum_par_1 );
29
30 ti.AddNewParagraph();
31
32 // modify style and apply them on next paragraph
33 ts.SetBorder( wxRP_BOTTOMBORDER, *wxRED, 0.1 );
34 ti.SetActiveTextStyle( ts );
35 ps.SetTextAlign( wxRP_RIGHTALIGN );
36 ti.SetActiveParagraphStyle( ps );
37
38 ti.AddText( lorem_ipsum_par_2 );
39
40 ti.AddNewParagraph();
41
42 // modify style and apply them on next paragraph
43 ts.SetBorder( wxRP_RIGHTBORDER, *wxRED, 0.1 );
44 ti.SetActiveTextStyle( ts );
45 ps.SetTextAlign( wxRP_LEFTALIGN );
46 ti.SetActiveParagraphStyle( ps );
47
48 ti.AddText( lorem_ipsum_par_3 );
49
50 m_Report.AddItem( ti );

```

At the first, a basic report style named `rs` is created at line 4. This object is used for drawing of light gray rectangle filling the text item's area. It is necessary to define both border and background color as can be seen at lines 5 a 6 because no page item can be filled without existing border. The borders can be specified by combination of binary flags `wxRP_LEFTBORDER`, `wxRP_RIGHTBORDER`, `wxRP_TOPBORDER`, `wxRP_BOTTOMBORDER` or simply `wxRP_ALLBORDER`. This general style object can be assigned to any type of page item by using `wxPageItem::SetStyle()` function.

Next, a text style object called `ts` is created at line 9. This style type can be applied just onto text items and allows definition of text font. This style can be assigned to the text item by using `wxReportTextItem::SetActiveTextStyle()` function as shown at line 25. From this point, any text value added to the text item uses this text style until changed to another one.

Finally, the paragraph style object is created at line 13 to specify various paragraph's properties like paragraph spacing or text alignment. Also, both text and paragraph style objects allow definition of their own borders. Borders defined by paragraph style are applied onto all following paragraphs added to parent text item while borders defined in text style object are applied onto single words. Paragraph style object can be assigned to parent text item by using

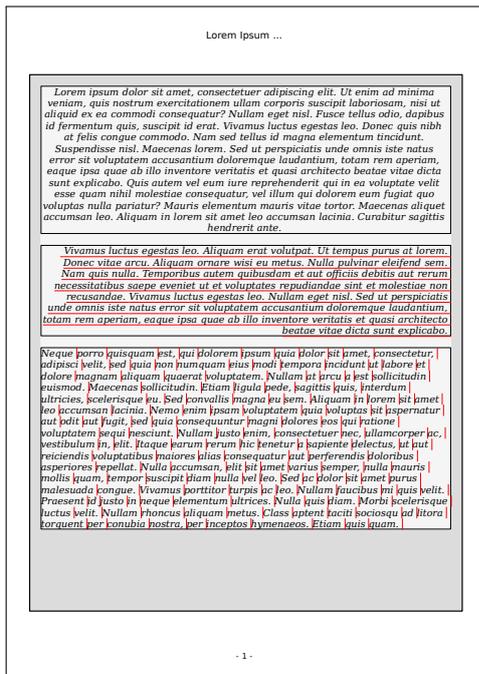


Fig. 6 styled text item

`wxReportTextItem::AddActiveParagraphStyle()` function. Fully styled text item is resulting from the sample code is show in Figure 6.

V. POSITIONING OF THE PAGE ITEMS

As mentioned in previous chapters, all page items can be positioned within custom page area or within the headers/footers areas by using absolute coordinates which are specified in *millimeters*. In addition to the absolute positioning, also predefined, automatically calculated position marks are available. These are:

- `wxRP_LEFT`,
- `wxRP_RIGHT`,
- `wxRP_TOP`,
- `wxRP_BOTTOM`,
- and `wxRP_CENTER`.

All position marks can be used for both vertical and horizontal positioning as shown at line 20 in Listing 8.

VI. PAGE ITEMS

In general, printed pages consist of set of page items with user-defined properties and styles. The following sub chapters describes all available item types in details.

A. Text items

Text items encapsulated by `wxReportTextItem` class can be used for inserting of single words or complex paragraphs as shown in Chapter IV.

There is also another important feature of the text item: it can display content of assigned variable updated at *run-time* automatically or on demand. It means that real displayed text

do not has to be "hard-coded" in source code, but can be read from program variables. Assignment of source variable can be done via `wxReportTextItem::AddVariable()` function which can take reference to short, int, long, float, double, char and `wxString` variables as its argument. This feature is ideal for definition of printed forms filled with calculated values, text phrases, etc. Values from assigned variables are converted to its textual representation automatically before printing/previewing or when requested by user via `wxReportDocument::RefreshVariables()` function.

B. Tables

Another very useful page item provided by `wxReportDocument` library is *table item*. As expected, the *tables* can be used for printing of data in tabular form. The table item is encapsulated by `wxReportTableItem` class which defines rich API suitable for handling its data, dimensions and style. The class members allow user to customize the table in many ways: if required, each table cell can be styled in completely different way, data can be inserted into the table per cell, per row and also per column. The table cell can contain single textual value or assigned variable used for filling the cell with desired value at run-time similarly to the text item as mentioned above (actually, the table cell *is* the text item so it provides similar functionality). Moreover, the tables can contain row or cell headers and are divided into several blocks fitting parent document page automatically when needed.

Now, let us demonstrate the table item's API. The example shows how to create simple table and how to fill it with values stored in array of integers. The sample code could be as follows:

Listing 9: Simple table item

```

1 // create array on integers
2 wxArrayInt columnValues;
3 for( size_t i = 0; i < 80; ++i ) columnValues.Add( i * 10
4 );
5 // define page style
6 wxReportPageStyle pgs;
7 pgs.SetWidth( 210 );
8 pgs.SetHeight( 297 );
9 pgs.SetMargins( 10, 10, 30, 30 );
10
11 m_Report.SetPageStyle( pgs );
12
13 // create table item and center it within document page
14 wxReportTableItem table;
15 table.SetPosition( wxRP_CENTER, 10 );
16 table.SetTextAlign( wxRP_LEFTALIGN );
17
18 // define default style used also for table headers
19 wxReportTextStyle cellsStyle;
20 cellsStyle.SetBorder( wxRP_BOTTOMBORDER );
21 table.SetCellStyle( cellsStyle );
22
23 // insert six columns with custom headers into the table
24 for( size_t c = 0; c < 6; ++c ) table.AddColumn(
25     columnValues, wxString::Format( "Header_%u", c ) );
26
27 // style specific table rows
28 cellsStyle.SetBorder( wxRP_NOBORDER );
29 for( size_t i = 0; i < 80; ++i ) table.SetCellStyleForRow(
30     cellsStyle, i );
31
32 // adjust table cells sizes automatically
33 table.SynchronizeCellsSizes();
34
35 // add the table to the document
36 m_Report.AddItem( table );

```

The most interesting aspect of this sample is the way how the table content is defined: the table columns are filled with values stored in array on integers as can be seen at line 24. Also, the

Header 0	Header 1	Header 2	Header 3	Header 4	Header 5	Header 0	Header 1	Header 2	Header 3	Header 4	Header 5
0	0	0	0	0	0	430	430	430	430	430	430
10	10	10	10	10	10	440	440	440	440	440	440
20	20	20	20	20	20	450	450	450	450	450	450
30	30	30	30	30	30	460	460	460	460	460	460
40	40	40	40	40	40	470	470	470	470	470	470
50	50	50	50	50	50	480	480	480	480	480	480
60	60	60	60	60	60	490	490	490	490	490	490
70	70	70	70	70	70	500	500	500	500	500	500
80	80	80	80	80	80	510	510	510	510	510	510
90	90	90	90	90	90	520	520	520	520	520	520
100	100	100	100	100	100	530	530	530	530	530	530
110	110	110	110	110	110	540	540	540	540	540	540
120	120	120	120	120	120	550	550	550	550	550	550
130	130	130	130	130	130	560	560	560	560	560	560
140	140	140	140	140	140	570	570	570	570	570	570
150	150	150	150	150	150	580	580	580	580	580	580
160	160	160	160	160	160	590	590	590	590	590	590
170	170	170	170	170	170	600	600	600	600	600	600
180	180	180	180	180	180	610	610	610	610	610	610
190	190	190	190	190	190	620	620	620	620	620	620
200	200	200	200	200	200	630	630	630	630	630	630
210	210	210	210	210	210	640	640	640	640	640	640
220	220	220	220	220	220	650	650	650	650	650	650
230	230	230	230	230	230	660	660	660	660	660	660
240	240	240	240	240	240	670	670	670	670	670	670
250	250	250	250	250	250	680	680	680	680	680	680
260	260	260	260	260	260	690	690	690	690	690	690
270	270	270	270	270	270	700	700	700	700	700	700
280	280	280	280	280	280	710	710	710	710	710	710
290	290	290	290	290	290	720	720	720	720	720	720
300	300	300	300	300	300	730	730	730	730	730	730
310	310	310	310	310	310	740	740	740	740	740	740
320	320	320	320	320	320	750	750	750	750	750	750
330	330	330	330	330	330	760	760	760	760	760	760
340	340	340	340	340	340	770	770	770	770	770	770
350	350	350	350	350	350	780	780	780	780	780	780
360	360	360	360	360	360	790	790	790	790	790	790
370	370	370	370	370	370						
380	380	380	380	380	380						
390	390	390	390	390	390						
400	400	400	400	400	400						
410	410	410	410	410	410						
420	420	420	420	420	420						

Fig. 7 simple table item spread over two document pages

column's header is defined there. In addition, two different styles are used with the table. The first one defined at line 18 is used for both table cells and the column headers while the second one defined at line 27 is assigned just to specific table rows. Notice, that the extent of the table exceeds document page dimensions so the table is divided into two pieces. The result of this sample code can be seen in Figure 7 which shows both printed pages.

C. Images

In addition to standard document elements like text item or table item mentioned in the previous chapters, the *wxReportDocument* library supports also set of static graphics items suitable for improvement of the document look.

The first item we are going to discuss here is *image item* encapsulated by *wxReportImageItem* class able to display images loaded from file system. The image item can be positioned similarly to text or table items within the custom page area od header/footer. It can be used in conjunction with dedicated style class called *wxReportImageStyle* allowing user to set image borders or margins. Also, PPI value can be customized via *wxReportImageItem::SetPPI()* API function so the image will be scaled in accordance to the set PPI value.

Let us to illustrate usage of *image item* on a simple example.

Listing 10: Simple image loaded from file system

```

1 // define page style
2 wxReportPageStyle pgs;
3 pgs.SetWidth( 210 );
4 pgs.SetHeight( 297 );
5 pgs.SetMargins( 10, 10, 30, 30 );
6
7 m_Report.SetPageStyle( pgs );
8
9 // define image style
10 wxReportImageStyle is;
11 is.SetBorder( wxRP_ALLBORDER, wxColour(100, 100, 100),
12             0.75 );
13 // create image item with custom PPI and apply custom
14 // style on it.
15 wxReportImageItem img;
16 img.SetPath( "../utb.png" );
17 img.SetPPI( 200 );
18 img.SetPosition( wxRP_CENTER, wxRP_CENTER );
19 img.SetStyle( is );
20 m_Report.AddItem( img );

```

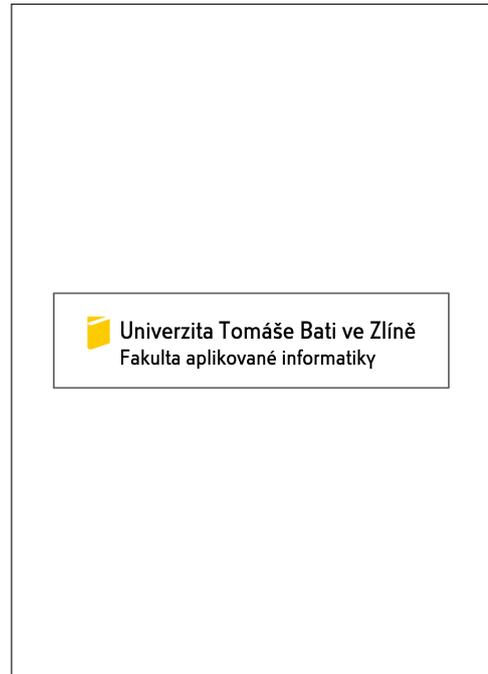


Fig. 8 centered image item

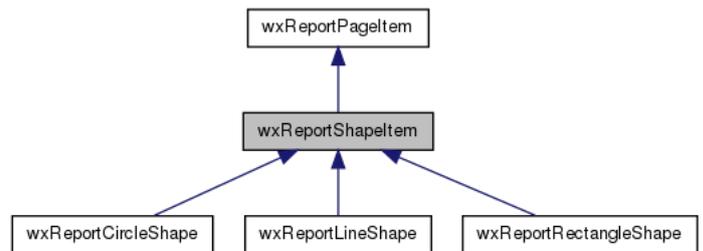


Fig. 9 hierarchy of shape items classes

The Listing 10 shows how to use custom image style defined at lines 10 and 11 and how to scale the source image by altering its PPI value as can be seen at line 16. Also, built-in position marks *wxRP_CENTER* are used for centering of the image within the main page area. Output of this source code can be seen in Figure 8.

D. Static graphics

Another way how to improve printed document's look is to use basic static graphic primitives like lines, rectangles and circles for highlighting of important parts of the document. For that purpose, the *wxReportDocument* library provides in addition to the *image item* also set of classes encapsulating these graphic objects which may be added to the document page. In contrast to the *image item* no position marks like *wxRP_LEFT* or *wxRP_CENTER* can be used for the placement - only absolute positioning is allowed there.

All static graphic primitives are inherited from common base class *wxReportShapeItem* defining API functions used for styling of the item. The shape base class allows definition on border/line color, line thickness and style and also definition of fill color if applicable.

Rectangles Probably the most common graphic primitive used in various printed documents and forms is *rectangle item*. This document item is encapsulated by `wxReportRectangleShape` and allows user to define its position, size, border and fill color. The usage of this class is straightforward and is illustrated in the following Listing 11.

Listing 11: Definition of rectangle shape

```

1 // define document page style
2 wxReportPageStyle pgs;
3 pgs.SetWidth( 210 );
4 pgs.SetHeight( 297 );
5 pgs.SetMargins( 10, 10, 30, 30 );
6
7 m_Report.SetPageStyle( pgs );
8
9 // create rectangle shae
10 wxReportRectangleShape rect;
11 rect.SetTopLeftCorner( 0, 0 );
12 rect.SetWidth( 190 );
13 rect.SetHeight( 237 );
14 rect.SetLineColor( *wxBLUE );
15 rect.SetFillColor( *wxGREEN );
16 rect.SetLineThickness( 4 );
17
18 // add the shape to the page
19 m_Report.AddItem( rect );

```

Circles Circles can be created and added to the document page by using `wxReportCircleShape` as shown in Listing 12. The *circle item* uses the same positioning and styling policy like the *rectangle item*.

Listing 12: Definition of circle shape

```

1 // define document page style
2 wxReportPageStyle pgs;
3 pgs.SetWidth( 210 );
4 pgs.SetHeight( 297 );
5 pgs.SetMargins( 10, 10, 30, 30 );
6
7 m_Report.SetPageStyle( pgs );
8
9 // create circle shape
10 wxReportCircleShape circle;
11 circle.SetCentreCoord( 95, 95 );
12 circle.SetRadius( 95 );
13 circle.SetLineColor( *wxBLUE );
14 circle.SetFillColor( *wxGREEN );
15 circle.SetLineThickness( 4 );
16
17 // add the shape to the page
18 m_Report.AddItem( circle );

```

Lines The last static graphic item supported by the library is *line shape*. As expected, the item allows users to define lines with specified starting and ending point, color and style. The class encapsulating this item is `wxReportLineShape` and can be used like demonstrated in Listing 13.

Listing 13: Definition of line shapes

```

1 // define page style
2 wxReportPageStyle pgs;
3 pgs.SetWidth( 210 );
4 pgs.SetHeight( 297 );
5 pgs.SetMargins( 10, 10, 30, 30 );
6
7 m_Report.SetPageStyle( pgs );
8
9 // create line shape and specify its style
10 wxReportLineShape line;
11 line.SetLineColor( *wxRED );
12 line.SetLineThickness( 2 );
13
14 // add a line leading from top-left to right-bottom page
    corner

```

```

15 line.SetPoints( wxRealPoint(0, 0),
16 pgs.GetSize() - wxRealPoint( pgs.GetLeftMargin() + pgs.
    GetRightMargin(),
17 pgs.GetTopMargin() + pgs.GetBottomMargin()
    ) );
18 m_Report.AddItem( line );
19
20 // add a line leading from top-right to left-bottom page
    corner
21 line.SetPoints( wxRealPoint(pgs.GetSize().x - pgs.
    GetLeftMargin() - pgs.GetRightMargin(), 0 ),
22 wxRealPoint(0, pgs.GetSize().y - pgs.GetTopMargin
    () - pgs.GetBottomMargin() ) );
23
24 m_Report.AddItem( line );

```

VII. CONCLUSION

As shown in the paper, `wxReportDocument` add-on to `wxWidgets` GUI toolkit can be regarded as fully functional, production-ready library helping users to define, print and even store various documents and forms easily. Notice that not all provided functionality is covered by this article. Not mentioned interesting features are for example an ability to define custom page items or to serialize/deserialize document content to/from XML file.

The library itself is published under `wxWidgets` License which means that there are no legal restrictions for using the library so it can be used for both open-source and closed-source projects freely. Thanks to that the library has been already used in two commercial project developed at our university which proved its maturity.

Source code of the library can be obtained from `wxCode` source repository [1] and can be build against `wxWidget 2.8.x` and `wxWidgets 3.0.x` GUI toolking by using any of supported compiler. In addition to the sources also Doxygen-based documentation [4] is available at the same location.

REFERENCES

- [1] `wxCode`. (2015). `wxReportDocument` Component. [Online]. Available: <http://wxcode.sourceforge.net/>
- [2] `wxWidgets` Website. (2015). [Online]. Available: <http://wxwidgets.org/>
- [3] `wxHtmlEasyPrinting Class Reference`, `wxWidgets` Website. (2015). [Online]. Available: http://docs.wxwidgets.org/3.0/classwx_html_easy_printing.html/af788066cba7ec33fe89bb66475729500
- [4] `Doxygen` Website. (2015). [Online]. Available: <http://www.stack.nl/~dimitri/doxygen/>
- [5] J. Smart, "Cross-platform GUI programming with `wxWidgets`", Upper Saddle River: Prentice-Hall, 2006.

Automation of modern marketing tools

Dagmara Bubel

Czestochowa University of Technology
Dabrowskiego 69, 42-200 Czestochowa, Poland

Abstract— The main aim of this publication is to identify possibilities provided to modern enterprises by the use of processes which are components of the system of marketing automation. The concept of marketing automation represents a completely new reality; it is a shift from communication that is based on mass distribution of a uniform message to communication that is realistically personalised, individual and fully automated. This is a completely new idea, a kind of coexistence in which sales and marketing departments closely cooperate with each other to achieve the best possible results. This is also a situation where marketing can confirm its contribution to the revenue generated by an enterprise. However, marketing automation also means huge analytical possibilities; it is real growth in the value of an enterprise, its added value represented by the set of knowledge acquired by the system about customers and all the processes performed in an enterprise, both marketing and sale processes. The introduction of a marketing automation system leads to a change in the existing model of functioning, not only of the marketing department but also a marketer. Everything that is provided by marketing automation, including cumulated unique knowledge about the customer, is also the critical marketing value of every modern enterprise.

Keywords— marketing automation, marketing tools, marketing communication, sales process.

I. INTRODUCTION

EVERY industry faces a range of specific challenges and actions. One of key, most important tasks is to build permanent, partner relations with customers that are based on the principles of mutuality. In the age of slow economic growth, winning a customer is an important process in each enterprise. This process becomes significant, if it makes it possible, by using its capabilities, to answer the following question: how to effectively win customers, then optimise the cost of this process, and finally persuade the customer to make a purchase. Universal access to the Internet resources and development of modern technologies, as well as constant changes of marketing models, make it more difficult to implement effective campaigns without the use of integrated systems of marketing automation.

This expanding range of possibilities was noticed already 10 years ago by marketers in the United States, leading to the emergence of processes that enabled monitoring of the behaviour of e-mail recipients beyond the mere opening of a

message. This is how processes known today as marketing automation were created. The main task of such systems is to monitor the behaviour of the recipient of a message on a particular website. Information obtained in this way can be used to identify the needs of the recipient and to specify which

elements of the offer are most interesting to potential customers. This information allows a message to be adapted to the needs of a particular potential customer, and if sent at the right time, it can strengthen the impact of the message. In modern world, the Internet has become an essential element of a consumer's existence, a place where products and services are searched for and where direct purchases are made.

In simple terms, marketing automation is technology that allows companies to improve marketing processes, better organise their tasks, fully automate applied strategies and precisely evaluate their effectiveness, and consequently to achieve a significant increase in ROMI (Return On Marketing Investment) [1]. The aim of such a system is to automate routine marketing tasks, and the functions of Marketing Automation include, among other things [2]:

- maintenance of databases about the existing and potential customers of the company as well as e-mailing lists,
- monitoring and analysing customer behaviour in mobile applications and behaviour of visitors to certain websites,
- segmentation of a company's potential customers by specific details, such as age, sex, place of residence or interests,
- monitoring the recipients of sent e-mails,
- management of B2B visits - identification of companies visiting a website,
- automated management of marketing campaigns.

A typical system of marketing automation provides a range of possibilities, both in terms of monitoring the behaviour of contacts on a website and reaction to such behaviour [3]. It is an answer to a new situation that occurred on the market as a result of a few factors related with both the model of customer behaviour and the way of reaching them with the offer. Marketing automation systems emerged as an expansion of e-mail marketing systems extending their capabilities to include the function of monitoring user behaviour on a website. This led to further development of these systems, which now only slightly resemble the first solutions from this field.

Nowadays, marketing automation is a system designed as a

tool for comprehensive support of marketing and sale activities in a company, with a focus on integration and synchronisation of processes taking place in these two areas. Lack of cooperation between marketing and sale is now one of the main reasons for failure of classical marketing. A change in the purchasing model, customers' taking control over this process, and a wide access to information from various sources - all these factors have had a huge impact on marketing effectiveness.

II. NEW CONDITIONS OF THE PROCESS OF COMMUNICATION IN THE ASPECT OF MARKETING AUTOMATION

Promotional activities of an enterprise undergo continuous changes, which is mostly determined by changing popularity of particular media. New media (the so called mobile and online media) show the biggest growth in the share in consumed media basket. For mobile media, the year on year growth in consumption is over 50% [4]. Data clearly shows an increasing role of new media [5]. Changes that are currently occurring have already affected activities undertaken by numerous enterprises in the area of marketing communication [6].

Enterprises' communication campaigns based on the concept of insideout no longer attract the attention of customers [7]. Advertising campaigns that used to be crucial for promotional campaigns show decreasing effectiveness today. Therefore, the answer to problems connected with decreasing level of effectiveness and acceptance of promotional activities of an enterprise is communication through new media. New media changed the position of stakeholders in the process of communication with an enterprise. Nowadays, it is the customer that initiates interactions with an enterprise.

In order to establish relations with the customer, an enterprise often has to offer valuable information in new media, sometimes even outside the main domain of its business activity. The customer usually finds valuable, trustworthy content using an internet browser. Therefore, an effective way to convey a message is to create value for the customer by providing interesting educational, entertainment or factual information, instead of sending promotional messages or messages intended to stimulate sales.

Communication activities focused on the implementation of a promotional function should be based on a very careful and in-depth analysis of the recipient of the message. That's why software for automating marketing campaigns is so popular [8]. The achievement of desired sales stimulation effects requires an in-depth knowledge about the customer, based not only on social and demographic data, but above all on behavioural information. Unfortunately, obtaining data about customers' communication or purchase behaviour is a task that takes time, IT and procedural solutions, not to mention the customer's consent to gathering information about their behaviour.

Today, enterprises have access to information and communication technologies, and there are no technical restraints of processing large collections of information.

Enterprises are equipped with tools such as marketing automation that enable them to generate a customized offer, but there is an obstacle on the way to fully use the marketing potential of technologies, which is the customer's willingness to provide access to their data to an enterprise. Therefore, the main barrier to the use of technology potential is access to data and the customer's consent to collecting, storing and processing of this data as well as making it available.

The functioning of marketing automation is presented in figure 1, and its main tasks can be divided into [9].

Outbound marketing - systems of marketing automation, complementing e-mail marketing systems, which have functions designed to distribute content to recipients from their database. This process, referred to as outbound marketing, can be implemented by sending e-mails, SMS or MMS messages, dynamic generation of advertisements or content of a website.

Inbound marketing — is a different strategy for winning new customers, which involves actions to ensure that the potential customer finds the offer he is interested in by himself. Marketing automation may be used both to generate websites and advertisements, and to suggest appropriate content in a blog or social media. The main role of marketing automation in this case is to provide accurate analytical data to better adapt the content being created to the needs of recipients.

Analysis of needs – in order to make the message reach their recipients, both in the case of inbound marketing and outbound marketing, it is necessary to gather appropriate data about potential customers. One of the main advantages of marketing automation is the possibility to collect a wide range of data about customer behaviour in the real time and to link it with information about each of potential customers in the database, which allows marketing to be customized to each recipient according to the one-to-one principle.

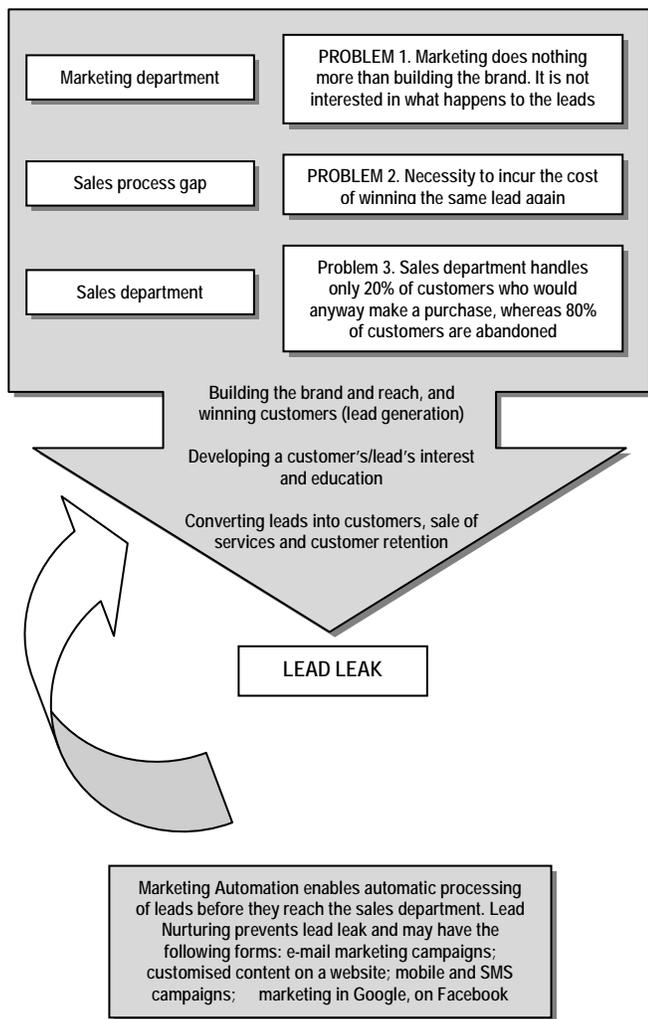


Figure 1. Functioning of marketing automation
Source: own work

In this discussion, it is necessary to mention lead generation, i.e. an activity that isn't directly connected with marketing automation, but is necessary for it to work properly. Systems of this type, although they are not originally designed to increase traffic on a website, require implementation of tools for increasing website traffic in an organisation. The main aim of lead generation is to gather e-mail addresses or other contact data, e.g. telephone numbers, of potential customers. This objective can be implemented in a number of ways, often incorrectly referred to by the common term "advertising." The process of obtaining contact data is mainly implemented by means of website forms that are completed during performing various activities, such as subscribing a newsletter, receiving a discount code, downloading special, otherwise unavailable documents. In terms of increasing traffic on a website and the number of people interested in completing forms, lead generation tools can be divided into [10]:

- SEO (Search Engine Optimization), i.e. positioning of a website in internet browsers,
- viral marketing - creating a video, picture, infographics or

internet memos shared by Internet users as something worth seeing,

- presence in social media - running a corporate website, establishing interactions with subscribers, providing valuable content,
- keeping a thematic blog connected with business activity.

Lead generation is often confused with demand generation due to the fact that it partially uses the same tools. However, there is a diametrical difference between these strategies in terms of the purpose of their use. Lead generation is intended to obtain the biggest possible number of contacts. To achieve that, it is necessary to use forms and offer some valuable content in exchange for leaving contact data. In contrast, demand generation consists in providing information about an offer to the widest possible group of receivers. To meet this condition, it is necessary to eliminate all barriers that can reduce the dissemination of information to potential customers, including all kinds of forms. Although lead generation and demand generation often use the same tools, these are strategies that practically cannot be applied at the same time [11].

III. USE OF MARKETING AUTOMATION TO SUPPORT A SALES PROCESS

Sales funnel is an illustration of a theoretical road travelled by an Internet user from entering a website to becoming a customer. The exact shape and number of stages of a sales funnel depend on a number of factors, such as industry, company, customer profile, product profile, sales and marketing targets, situation in the environment of an enterprise, available offer, applied marketing tools etc. Therefore, sales funnels vary across companies. Often, there are a few funnels within one company. Below are a few key elements for the functioning of a funnel [12]:

- attracting visitors - preliminary phase for a sales funnel. Leads can be obtained through inbound and outbound marketing;
- Conversion of visitors into leads - leaving contact data;
- Qualification of leads - assessing which of the leads obtained are suitable for further actions, as a substantial part of leads provide false data;
- Conversion of leads into MQL (Marketing Qualified Leads) — MQLs are leads that have been qualified as requiring marketing campaigns. The conversion takes place by subscribing a newsletter, downloading documents or participating in a webinar. This point opens the door to the operation of the system of marketing automation.
- Lead nurturing - aimed at converting leads into purchase leads. It usually takes form of a cyclic mailing containing valuable content, but more and more often it is a dynamic and multi-channel process.
- Conversion of a lead into SQL and transferring it to a sales department - conversion of a lead into SQL (Sales Qualified Lead) is the aim of creating a sales funnel. The system of marketing automation is invaluablely helpful here, as it is able to automatically recognise a real conversion of

leads and transfer them to sales departments without the interference of marketing staff.

- Further activities with a new customer - depending on the strategy, the funnel can be stopped at this point or continued, e.g. to sell more products to the customer. This point is not permanent, and may not be present in certain industries. It can also be implemented in several separate sales funnels.

Marketing automation systems work in each of the points above, but the most important area of their operation is the space between the marketing department and sales department. In the classical sales model, there are no specified patterns of activities that should be taken at this stage. After generating leads, marketing departments leave them unattended or send all leads directly to sales departments. This is not an optimal solution, as it causes leads to leak excessively, results in a gradual decrease of the contact database, overload of sales departments, necessity to obtain the same lead anew, general chaos and lack of cooperation between marketing and sales.

Lack of communication between marketing and sales is mainly the effect of a business model in which these departments function separately. In the case of traditional marketing tools, this was a reasonable step, but today, in internet marketing, it significantly limits the possibilities of acting. Nowadays, we can see a need to fill a lead gap - the gap between generated leads and the sales department. Activity in this area is characterised by certain limitations resulting from changes in customers' purchase models.

In the age of the Internet, it is usually assumed that the control of the purchase process lies with a potential customer. The customer decides whose offer to read, what content to receive, which one to accept and which one to reject. The modern customer is less prone to be influenced by advertisement, and relies more on other people's opinions, documented articles in trade press and knowledge shared in specialised portals. He absorbs messages both in text and audiovisual forms. He is willing to establish contacts with other users with similar problems, but is sceptical about interaction with sales departments of companies whose offer he views. For that reason, an effective activity in the area of lead gap requires providing the customer with valuable informational content and avoiding activities that have the character of an offer.

The process of impacting a potential customer by providing him with high quality factual content is called lead nurturing. Lead nurturing is one of the most important methods for using marketing automation systems to increase the number of potential customers. Lead nurturing is a new marketing term connected with conducting marketing campaigns designed to prepare a potential customer to make a purchase. It is most often used in the area of B2B sale of products and services. The aim of a lead nurturing project is to provide potential customers with knowledge and information necessary for the sales department to conduct a sales campaign [13].

The main advantage of lead nurturing is constant control over the interaction between a lead and the company. To

achieve that, it is necessary to establish the actual demand of a lead for information, find out what are his interests or requirements, and to react by providing him with valuable content, especially in the face of information chaos [14].

In its most simple form, lead nurturing consists in ensuring educational care to a contact by cyclically sending him messages, e.g. in the form of a newsletter. More advanced forms may use varied channels to reach a lead, including online webinars, social media, e-mail marketing, SMS messages and dynamically generated websites. It is also possible to use all these channels interchangeably or at the same time. A frequently used practice is to support the process of lead nurturing with telephone conversations.

Lead nurturing projects have a few objectives [15]:

- conversion of leads - the main and fundamental aim of lead nurturing is to impact the conversion of a lead to next stages of a purchase process, i.e. graphically speaking to walk a lead through a sales funnel.
- maintaining contact with a potential customer - which is necessary to sell a product or service. Thanks to lead nurturing, the contact between the company and potential customers is not broken off. As a result, it is not necessary to regain the same leads, e.g. through advertising.
- Provision of key information - thanks to lead nurturing, a company can take partial control over the information reaching its potential customer. It can include content that supports conversion.
- Image creation - by offering valuable, useful content to leads, their creator is perceived as a specialist in a given industry.
- Indicating appropriate leads to sales departments - thanks to modules of assessment of potential customers in terms of their readiness to make a purchase, it is possible to automatically transfer leads that underwent conversion into SQL in the real time. Such activity significantly shortens the response time of a sales department, increases sales, and frees sales specialists of the necessity to contact with "cold leads".
- Reduction of operating expenses - thanks to automatic distribution of contents and automation of sales funnels it is possible to reduce costs generated by marketing and sales, both in terms of savings on campaigns and reduced need for staff

IV. USE OF MARKETING AUTOMATION IN THE LIGHT OF RESEARCH

Marketing automation systems have evolved for years, going a long way from their original form of expanded systems for e-mail marketing. Nowadays, they are tools that collect data from numerous sources and have a multi-channel impact on potential customers. Over more than ten years of the existence of this type of systems, a range of ways to use them to support marketing and sales have been developed. This goes far beyond the original purpose of marketing automation, i.e. guiding leads inside sales funnels.

Research showed that implementation of marketing

automation leads to increased frequency of mailings and at the same time improvement of their effects. 58% of those surveyed confirmed sending up to 20 e-mails monthly. 47% of those surveyed send fewer than 10 e-mails per month. The latter group contains the biggest number of companies that do not use any technologies intended for e-mail marketing mailings. They account for 43% of companies [16].

Another interesting conclusion of the survey is the fact that the use of traditional e-mail marketing platforms does not affect the frequency of mailings. 61% of companies that use such platforms send fewer than 10 e-mails monthly. For comparison, in the case of companies using marketing automation only 31% send fewer than 5 e-mails per month. Moreover, this percentage falls to 20% when a company uses marketing automation longer than one year [17].

Based on research we can indicate three main concerns connected with implementation of marketing and sales automation [18]:

- Has the company got appropriately qualified staff to use new tools?
- Can Marketing Automation bring real benefits?
- Can the enterprise afford to implement Marketing Automation?

The survey shows that [19]:

- 65% of companies expect that ROI (return on investment) will become the most important indicator of the measurement of their business results over the next three to five years, whereas only 46% are prepared for that;
- 79% of marketing departments have not developed a method for turning leads and potential customers into customers who pay;
- 63% of marketers refer every, even weak, lead directly to the sales department, although maximum 30% of such leads qualify for that;
- 91% of company owners cut marketing budgets in the first place.

Nowadays, we can notice a change in purchase models, which is directly connected with growing Internet penetration, while research shows that [20].

- A purchaser wants to control as much as 70% of the purchasing process
- 90% of consumers do online research before the purchase
- This shows that there is a gap in the sales process (lead leak), as the purchaser makes a decision to purchase a product much earlier. The gap in the sales process and how it is filled by the Marketing Automation system is illustrated in figure 2.

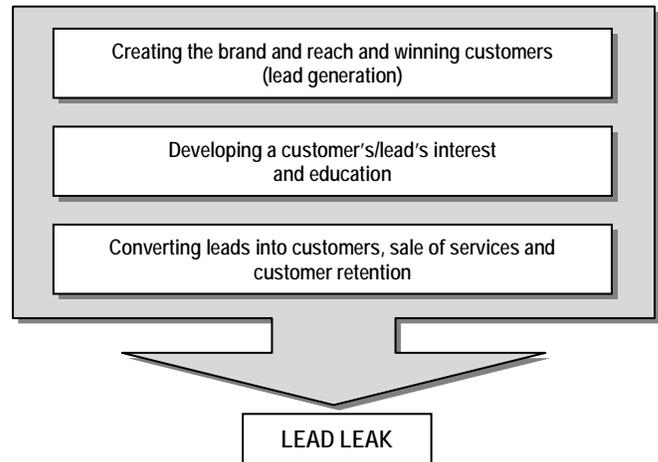


Fig. 2 Impact of the marketing automation system on the gap in the sales process

Source: own work

The survey shows that more and more marketers perceive advantages of marketing automation, and this system is currently used by 56% of B2B marketers. 19% consider starting to use this system. 17% claim that they are aware of the advantages of this solution, but they are not using it yet. Of companies that implemented the MA system, 78% have been using this solution for over a year, whereas 29% - for less than 12 months. More than 15% of companies have been using this system for over 5 years. An interesting issue is the period of time between the purchase of the system and its full use. The survey shows that the majority of marketers (81%) implemented the system within 6 months [21].

When asked about the advantages of using this system, marketers mention: more leads and their better quality (68%), higher effectiveness of marketing campaigns (51%) and better conversion rates (42%). Better rates of consumer involvement were indicated by 33% of respondents, whereas 35% noticed the advantage of better possibilities of consumer targeting. For 32%, an important thing was shortening the sales cycle, whereas for 25% - bigger traffic on the website. The most important functionalities of the Marketing Automation system were, as rated by those surveyed: lead nurturing (52%), integration with CRM, social media and mobile platforms (49%), analysis and reports (47%), lead scoring (39%), management of campaigns (38%), e-mail marketing (37%) and monitoring of consumer activity (38%). Those surveyed mentioned also obtaining leads, monitoring visits on websites, segmentation, content publication, monitoring social media and indicators connected with consumer involvement. Among the difficulties encountered on the way to make a full use of Marketing Automation, marketers indicated, among other things: budgetary constraints (39%) and lack of qualified staff (36%) [22].

Functionalities, benefits and difficulties connected with the system of marketing automation are presented in figure 3.

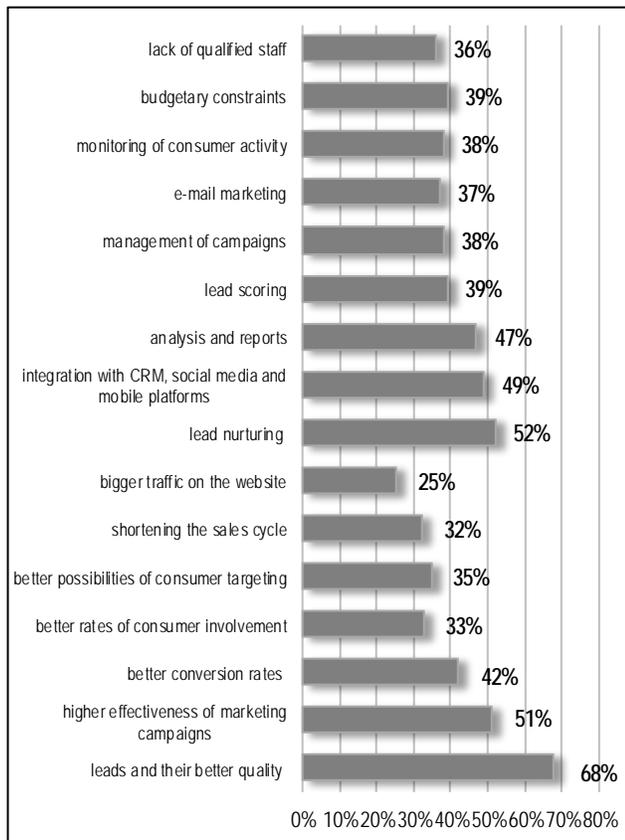


Fig. 3. Functionalities, benefits and difficulties connected with the system of marketing automation.

Source: own work based on literature.

61% of the companies surveyed use the Marketing Automation system to some extent, whereas as many as 37% don't use it at the moment, but they plan to introduce this solution in the near future. Only 10% of the enterprises surveyed do not plan at all to use automation. 86% of marketers using Marketing Automation claimed that this system helped them to achieve important objectives, such as:

- sales growth – 49%,
- obtaining a bigger number of leads – 49%,
- better lead nurturing – 45%,
- increased customer involvement – 39%,
- increased effectiveness of marketing – 35%,
- better measurability of the effectiveness of actions – 27%,
- more precise focus of the campaign on a particular group of customers – 28%,
- better coordination of marketing and sales departments – 30% [23]

87% of those surveyed think that marketing automation is important for their comprehensive programme of marketing campaigns, with 66% regarding it as "very important" for the success of the strategy adopted. Over half of the companies surveyed (55%) plan to increase their Marketing Automation

budget within the next year, and 46% intend to keep it at the current level [24.]

Reports from surveys inform that 65% of directors of companies expect that the return on investment in marketing (ROI) will become the most important measure of their activities within the next three to five years, but only 43% feel that they are prepared for that. According to the majority of company owners (87%), the marketing department does not create the value of the company, does not take direct responsibility for revenue and usually features in financial reports as a component of cost centres. The negative standing of marketing departments in companies has slightly improved in recent years as a result of the increasing role of the Internet in the functioning of companies and the change of sales models, which led to specialist positions created in companies that require analytical knowledge and skills [25].

V. SUMMARY

The emergence of marketing automation is the result of changes of the conditions in managing marketing communication. Processes taking place in new media had the biggest impact on marketing communication. The role of the recipient in marketing communication is growing. Communication of an enterprise no longer fulfils only the function of promotional mouthpiece, because the recipient more and more actively initiates the process of communication with an enterprise. The effectiveness and efficiency of an enterprise's marketing campaigns is growing only if the recipient receives interesting, reliable and trustworthy content. Marketing automation, thanks to its possibilities, can be the most important solution in modern marketing. This thesis seems to be confirmed by the results for the whole marketing automation market, that grows by several dozen percent year by year.

Marketing automation perfectly fits the latest marketing trends, such as the use of Big Data, one-to-one communication or content marketing. However, its most important advantage is the fact that in the age of information chaos marketing automation can adapt the message to an individual recipient. Undoubtedly, modern functioning in the information society is connected with increasing amount of content coming from all directions to the Internet users. Thus, without adjusting the message to the current needs of recipients it is not possible to attract their attention. Other systems for managing relations with customers, despite their obvious advantages, have a certain fault connected with their construction - they are unable to react in the real time. The assumptions included in the process of marketing automation definitely modernise campaigns, which enables the achievement of much higher return on investment compared with using classical marketing tools. The introduction of the system of marketing automation leads to a change of the existing model of functioning not only in the marketing department but also the marketer. After all everything that marketing automation provides, including cumulated unique knowledge about the customer, is also the critical marketing value of every modern enterprise.

REFERENCES

- [1] J. Steinbach, M. Krisch, H. Harguth. „Marketing-Automation: integrierte Technologie einsetzen“, *Helpvertising*. Springer Fachmedien Wiesbaden, 2015, pp. 33-37.
- [2] A. Englbrecht, H. Hippner, K. D. Wilde. „Marketing Automation—Grundlagen des Kampagnenmanagements“, *IT-Systeme im CRM*, Gabler Verlag, 2004, pp. 333-372.
- [3] K. Plehwe, „Marketing-Automation und Kampagnenmanagement – Moderne Instrumente für den Erfolg des Dialogmarketing“, *Das Mailing*, Gabler Verlag, 2002, pp 33-43.
- [4] J. Steinbach, Jan, M. Krisch, H. Harguth. „Eine neue Philosophie im Marketing: Helpvertising statt Advertising“ *Helpvertising*, Springer Fachmedien, Wiesbaden, 2015, pp. 9-22.
- [5] D. Jelonek, „Przewaga konkurencyjna e-przedsiębiorstwa“, *Ekonomika i Organizacja Przedsiębiorstwa* 3 (2003), pp. 26-38.
- [6] B. Nogalski, G. Wejer, *Strategie marketingowe dla sektora MSP w przededniu wejścia Polski do Unii Europejskiej*, Prace Naukowe Instytutu Organizacji i Zarządzania Politechniki Wrocławskiej, Konferencje 22.71 (2001), pp 157-167.
- [7] T. Schwarz, „E-Mail-Marketing“, *Digitales Dialogmarketing*, Springer Fachmedien Wiesbaden, 2014, pp. 411-429.
- [8] J. Steinbach, M. Krisch, H. Harguth. „Die Entwicklung von hilfreichen Content“, *Helpvertising*, Springer Fachmedien Wiesbaden, 2015, pp. 23-32.
- [9] Ch.-M. Geiger, „Die Facetten der Adresse-Adressen- und Listmanagement“, *Digitales Dialogmarketing*. Springer Fachmedien Wiesbaden, 2014, pp. 303-325.
- [10] S. Koch, „IT-Unterstützung von Prozessen“, *Einführung in das Management von Geschäftsprozessen*, Springer Berlin Heidelberg, 2015, pp. 253-280.
- [11] A. Hermann-Ruess, „Der unternehmerische Mehrwert von Webinaren“, *Das gute Webinar*. Springer Fachmedien Wiesbaden, 2014, pp. 185-194.
- [12] Ch. Schawel, F. Billing. „Sales-Funnel-Analyse“, *Top 100 Management Tools*, Gabler Verlag, 2014, pp. 222-224.
- [13] T. Hermann, „B2B-Inbound-Marketing – Aktive Interessenten als Kunden gewinnen“ *Marketing Review* St. Gallen 26.6, 2009, pp. 31-36.
- [14] A. Brzozowska, A. Nowakowska. *Rola platform informatycznych w polityce informacyjnej korporacji*, Prace i Materiały Wydziału Zarządzania Uniwersytetu Gdańskiego 1, 2009, pp. 53-58.
- [15] Steinbach, Jan, Michael Krisch, and Horst Harguth. "Eine neue Philosophie im Marketing: Helpvertising statt Advertising." *Helpvertising*. Springer Fachmedien Wiesbaden, 2015. 9-22.
- [16] M. Giordano, J. Hummel, eds. *Mobile Business: Vom Geschäftsmodell zum Geschäftserfolg—Mit Fallbeispielen zu Mobile Marketing, mobilen Portalen und Content-Anbietern*, Springer-Verlag, 2015.
- [17] J. Steinbach, M. Krisch, H. Harguth. „Erfolge messen und sichtbar machen“ *Helpvertising*, Springer Fachmedien Wiesbaden, 2015, pp. 39-45.
- [18] D. Spiegelberg, „Erfolgspotenziale durch Vernetzung“, *Enterprise Marketing Management*, Springer Fachmedien Wiesbaden, 2013, pp. 63-70.
- [19] J. Zirke, A. Wiersgalla, *Abbildung von B2B Prozessen in indirekten*, Praxis des Customer Relationship Management: Branchenlösungen und Erfahrungsberichte (2013), p. 164.
- [20] G. Heinemann, „Online-Handel der Zukunft.“ *Der neue Online-Handel*. Springer Fachmedien Wiesbaden, 2015, pp. 1-32].
- [21] M. Giordano, J. Hummel, eds. *Mobile Business: Vom Geschäftsmodell zum Geschäftserfolg—Mit Fallbeispielen zu Mobile Marketing, mobilen Portalen und Content-Anbietern*. Springer-Verlag, 2015.
- [22] T. Theuring, E. *Online-Marketing: Grundlagen, Modell und Fallstudie für Versicherungsunternehmen*, Springer-Verlag, 2013.
- [23] O. Rengelshausen, *Online-Marketing in deutschen Unternehmen: Einsatz—Akzeptanz—Wirkungen*, Springer-Verlag, 2013.
- [24] H. Gräf, *Online Marketing: Endkundenbearbeitung auf elektronischen Märkten*. Springer-Verlag, 2013.
- [25] J. Link, D. Tiedtke, eds. *Erfolgreiche Praxisbeispiele im Online Marketing: Strategien und Erfahrungen aus unterschiedlichen Branchen*. Springer-Verlag, 2013.

System for Professionals – monitoring employers’ demands for key competences in Wielkopolska

M. Szafrąński, M. Goliński

Abstract – The article presents a proposal for activities aimed at acceleration of access to information regarding key competences. It gives justification of why key competences are currently important for the effectiveness and efficiency of enterprises. System for Professionals, operating in Wielkopolska since 2013 is characterized. It was created and is being developed in order to monitor competences, especially key ones. An integral element of the system is the IT tool, facilitating data and information preparation for users, who are at the same time subjects of the System for Professionals. Sample results of monitoring carried out within the system are presented. They concern the needs of entrepreneurs in the Wielkopolska voivodeship for key competences. The survey included 918 enterprises registered in the system.

Keywords – key competences, IT system, skills, acceleration

I. INTRODUCTION

ACCCELERATION of activities aimed at achieving objectives of individuals must be accompanied by acceleration of creating knowledge resources in enterprises. This problem is discussed by M. Szafrąński [1]. The faster the environment of enterprises changes, the faster adaptation activities must take place in those enterprises. In the knowledge-based economy, one of the key factors for effective and efficient functioning on the market is knowledge. As a resource at the origin of processes in an enterprise, it should respond in a dynamic and continuous way to the changing needs resulting from targets. In workstations, it is necessary to increase knowledge, especially professional skills, in a continuous way. It is equally important, though, for the employees to possess well-developed key competences. Improvement of methods of monitoring the needs for key competences in businesses results in shortening the access time to information about gaps in these competences. Consequently, both businesses and institutions responsible for educational processes may react faster to minimize those gaps both in educational systems and in the labour market. One solution supporting acceleration of access to information about employers’ needs for competences, including key competences, is the System for Professionals, operating in Wielkopolska since 2013. This article presents sample survey results on monitoring the skills which are part of key competences.

II. KEY COMPETENCES AS AN ELEMENT OF KNOWLEDGE

Knowledge and skills are often perceived as separate elements, especially in the pedagogical perspective – in education. For example, European Qualifications Framework, in accordance with the recommendation of the European Parliament and the Council of the EU of 23 April 2008, divides learning outcomes [2, p.13] into knowledge, skills and competences (personal and social). Similarly, skills and knowledge are treated as separate categories in Dublin descriptors [3, p.27]. Management science, on the other hand, as well as some other sciences (e.g. praxeology), describe skills as a category of knowledge. In some sources [4, pp.31-33], [5, p.12], knowledge is divided into four different categories:

- 1) **know-what** – operational; it is a base for ordinary, everyday work; easy to describe with words and easy to transfer;
- 2) **know-how** – operational; hidden in human mind; connected with experience of how something is done and how this ‘something’ operates; difficult to describe with the language of signs; obtained through personal experience; may be identified with **skills**;
- 3) **know-why** – including awareness of goals defined in a business and justification for achieving them; identified with the awareness of changes taking place in the surroundings, knowledge of the context of actions undertaken;
- 4) **know-who** – knowledge on who is who and what knowledge they possess both in the surroundings and inside a business; knowledge on the role and status of subjects of activities within and outside an organisation.

G. Probst, S. Raub and K. Rombards treat skills as a component of knowledge. They suggest understanding knowledge as ‘all information and skills possessed by the subject of actions’ [6, p.9]. This paper also treats skills as a category of knowledge.

While analyzing literature, a great variety in classifying skills may be noticed. Skills are most often grouped in sets called competences. E. Kolanowska [7, pp. 321-322] undertook to organize various categories into competences, and although she mentions several, they should be treated as a set of notions rather than a classification. Detailed classifications of professional skills (e.g. [8]) are available, but with non-professional ones, there is a great variety when giving them names. This makes communication difficult between parties interested in information exchange

concerning skills. It concerns especially communication between employers who look for candidates possessing a certain set of skills or – putting it more widely – non-professional competences, and candidates themselves.

A special set is the set of key competences that are essential both socially and for most entrepreneurs. Literature of the subject describes this group of skills in detail. Assuming recommendations of the European Parliament and the Council of the EU, the following can be listed [9]:

- 1) communication in the mother tongue,
- 2) communication in foreign languages,
- 3) mathematical competence and basic competences in science and technology,
- 4) digital competence,
- 5) learning to learn,
- 6) social and civic competences,
- 7) sense of initiative and entrepreneurship,
- 8) cultural awareness and expression.

These competences may be divided into two sub-groups: traditional and horizontal ones [10]. Traditional competences include the first four from the above list, sometimes adding literacy [10], the latter four create the sub-group of key competences [10], [11].

III. SYSTEM FOR PROFESSIONALS – ORGANIZATIONAL-TECHNICAL INNOVATION FOR MONITORING COMPETENCES IN THE LABOUR MARKET

A motivation to carry out work in the System for Professionals was the AWT[®] programme [12]. AWT[®] is a programme, started in 2006 at Poznań University of Technology, called Technical Knowledge Accelerator (Akcelerator Wiedzy Technicznej[®]), aiming at increasing effectiveness and efficiency of activities shaping relationships between the educational systems and the labour market. Assumptions of designing the future are realized through acceleration. Innovative, unusual solutions are frequently chosen, since they are often more effective and more economical compared to classical, less effective ones, but ones with a lower risk of failure. Focus research and individual surveys aimed at analyzing the needs of employers showed that there is a lack of up-to-date, detailed and reliable information on competences and qualifications of candidates for jobs in enterprises. The needs of employers were the reason why goals realized in the System for Professionals are being developed. However, the System concentrates on managing information regarding the needs of the labour market. An innovative vision on solving the existing problems was essential for the creation and development of the system. Innovations are mainly of organizational character and are integrated networked [13, p. 37]. The digital solution – the central tool of the System for Professionals – developed and adapted every functionality in cooperation with its users, taking into consideration mutual relations in the labour market, and even in education. Cooperation of numerous subjects, frequently of varied character, is nowadays the basis for success of the innovation [14, 15]. Each of the modules of the system serves to realize a separate function, is evaluated and improved in each design and implementation iteration – agile software development [16]. Put together, the modules create a system of information exchange within the same databases, the same

authorization and authentication system. Easy access to information ensures a cohesive interface. The structure of the software (source code) and the composition of databases allow developing the tool through functional modifications and day-to-day improvement of the interface. Due to a large scope of tasks realized in the System and numerous groups of users, the system is composed of modules. In the latest version of the System, released in 2015, the following functionalities can be listed:

Module for the entrepreneur – the most important functionality of the system – enables to define precisely requirements for an employee in a particular position. The structure of the user interface gives an intuitive solution to create an offer for an employee, apprentice or trainee. The system makes it possible to describe the profile of competences, and thus plays the role of a mobile recruitment system. Free registration and use of the System help minimize costs and shorten the time – two significant parameters of recruitment process evaluation [17, 18].

Module for students/graduates is a base describing the potential of vocational schools students. The potential employee presents their competence profile, which is automatically compared against job offers in the base. The module provides for the evaluation of employee's competences based on 360° degree method [19].

Module of career counselling – enables planning the educational path and support in professional development. Referring to the structure of job description – built in a hierarchy from basic skills through competences and qualifications – the module may be used for the description and evaluation of a single job position [20] or to characterize human capital in the organization as a whole [21].

Module of manager of practical training – the module allows facilitating the process of managing trainings at the employer's site and to organize forms of employment other than full-time/partial-time, expected by employers [22].

Module of trainings – training companies publish their offers of courses and trainings in response to the needs of the labour market and complement the formal – school education.

E-learning module – along the System for Professionals there is a module of distance learning. It is a form of integrated education and self-education system, which together with 'anticipatory' vocational practice takes the form of triplex education – as a developed form of dual education [23].

Analytical module – allows generating real-time reports and bipartite analyses on the labour market and education in the Wielkopolska region; it also provides complex reports already generated and commented.

A dynamic growth in the number of the system users and positive opinions of institutions evaluating the System give the right to claim that it is a useful tool, bringing expected benefits.

IV. EXAMPLES OF MONITORING KEY COMPETENCES WITH THE USE OF THE SYSTEM FOR PROFESSIONALS

A. *The significance of key competences*

System for Professionals makes it possible to create numerous analysis sections. For example, significance of a competence from the perspective of employers can be

examined. Due to priorities accepted for the Wielkopolska region, the system currently gives the possibility to examine the employers' demand for skills obtained on the level of secondary vocational level only. According to Central Statistical Office of Poland, in the year 2012, there were as many as 8.56m out of 15.59m of economically active employees (i.e. nearly 55%) with secondary vocational education, post-secondary education or vocational education. In spite of this, entrepreneurs find a deficit of employees with professional skills. A similar situation can be observed in the region concerned. Therefore as early as in 2008, a bill [24] was passed by the Board of the Region of Wielkopolska. It defined the frames of vocational education development. Formal vocational education is perceived by enterprises as an element of preventive action, which M. Szafrński described in the context of vocational apprenticeship [25]. Adaptation of education towards acceleration of providing graduates with **professional skills** is one of the priorities of the educational system. However, employers expect candidates for jobs, apprenticeships and trainings to show a proper level of **skills that constitute key competences**. According to the representatives of employers who were interviewed, these skills accelerate employees' adaptation to work after hiring them [1]. Current scope of implementation of the System for Professionals assumed the classification of professions and skills as given by the Act of Law of Minister of Education on the curriculum for vocational education [8] (state law).

In the System for Professionals it is entrepreneurs who publish their offers of job, apprenticeship and training. These offers are received mainly by graduates of technical schools and vocational schools, but also by students of these schools (in case of apprenticeship and training). The offers include general information, indications of vocational skills and key competences expected by employers. As a result of the classification accepted [8], all skills are divided in the system into three sets:

- 1) skills common for all professions,
- 2) skills common for individual areas of education,
- 3) professional skills.

Skills common for all professions are in fact skills that constitute key competences. They are skills that – from the perspective of the Polish vocational education system – all graduates of vocational schools should possess. Therefore, employees with vocational education should also possess them. Monitoring offers published in the system by employers gives information as to which skills are most important from their point of view, and how demand for these skills evolves over time. It concerns also skills that constitute key competences, called 'common' in the system, in reference to all professions included in the classification of professions published by the Ministry of Education (MEN).

More than two years after the implementation of the system (January 2013 to April 2015), the number of common skills (connected to key competences) in the system is greater than the number of skills in the classification prepared by MEN, which was used as the starting point. A reason for this is also the language used in enterprises, substantially different from that of the curriculum, formalized to a large extent. Most employers reject partially the language of the curriculum, and since the 'life' of the system depends greatly on the

entrepreneurs' activity, the skills classification in the system evolves, showing imperfections of adapting the official systems of skills descriptions to practice. System for Professionals facilitates monitoring linguistic differences and similarities of two different environments (education and economy). On the one hand, it introduces phrases demanded by employers, and on the other hand, it helps keep the order of definitions, systemizing the language proposed by employers. Employees use most often colloquial language, which is imprecise, ambiguous, the use of which results in blurred notions. To use a completely deformed language in the system of information exchange would inhibit communication between parties participating in the information exchange. M. Szafrński [1] described the problem of ambiguity of phenomena connected with knowledge management defined by employers. It was manifested during research between 2010-2012, while creating the system in the part led by the author.

Although this article discusses key competences, and research results presented below concern skills related to them, it is necessary to add that professional skills included in the System for Professionals are those attributed to certain professions. Common skills in individual areas of education are connected to the area of education described in detail in the Act of Law on the curriculum in vocational education [8].

The basis for research on the demand for key competences in enterprises of the Wielkopolska region are data from the System for Professionals. On 10 April 2015 there were 918 enterprises registered in the system – they are the research sample. These enterprises, publishing offers of job, apprenticeship and training between 1 January 2013 and 10 April 2015, indicated demand for 112 common skills.

Monitoring shows that enterprises which use the system, indicate the demand for common skills in accordance with the Pareto-Lorenz principle. 70% of all indications of common skills occurring in offers translate to about 21% of skills, i.e. only 24 out of 112 mentioned in the offers as desired, in other words – necessary. **Indication** of a given skill means that the employer indicated it as expected from the candidate in the offer of job, apprenticeship or training. The offer may indicate many skills. The total number of indications of a given skill in the period concerned means that this skill was indicated in all offers that many times. Based on the above observation it may be stated that for enterprises which use the system, nearly 80% of common skills indicated at least once are unimportant or of marginal importance. The Pareto-Lorenz principle also applies to SMEs (70% of indications of 23 skills, constituting 21% of all 109 common skills indicated by SMEs) and large enterprises (70% of indications of 26 skills, constituting 26% of all 99 common skills indicated by large enterprises) regarded separately.

Even these basic results are a source of valuable information not only for enterprises, but also for institutions responsible for the educational policy and educational processes in the region. This information may translate into planning actions which would accelerate the increase of the level of key competences, essential for enterprises, and thus for the whole economy.

Research carried out suggests that only between 7 and 10 of common skills indicated in the System for

Professionals are those which make up 30% of all employers' indications (table 1).

Research up to date shows common skills for professions most often indicated by employers in the system. They are:

- 1) respect the rules of behaviour and ethics (4.5% of indications),
- 2) cooperate in a team (4.1%),
- 3) respect confidentiality (3.7%),
- 4) be creative and consistent in task realization (3.6%),
- 5) be able to cope with stress (3.4%),
- 6) be open to changes (3.4%),
- 7) foresee results of action taken (3.4%),
- 8) update knowledge and develop professional skills (3.3%).

These eight skills indicated in employers' offers as required were 29.4% of all 4078 indications of common skills registered.

Table 1. Relationships between common skills most often indicated and all common skills indicated in the System for Professionals: CS–common skills, SP–System for Professionals.

Period	All CS indicated at east once in SP	# of CS in all offers of job, apprenticeship an training in SP	About 30% of all indications of CS in SP (30%*(3))	Skills indicated in 30%	
				Number of skills from the list where all skills are ordered from the most frequently indicated to the least frequently indicated	% of skills indicated [5] / [2]
[1]	[2]	[3]	[4]	[5]	[6]
SMEs					
01.2013-06.2014	69	2622	863	8	12%
11.2013-04.2015	10 9	2834	919	9	8%
01.2013-04.2015	10 9	3470	1041	8	7%
LARGE ENTERPRISES					
01.2013-06.2014	61	258	83	6	10%
11.2013-04.2015	99	481	150	10	10%
01.2013-04.2015	99	525	161	9	9%
ALL ENTERPRISES					
01.2013-06.2014	69	2864	853	7	10%
11.2013-04.2015	11 2	3315	1048	9	8%
01.2013-04.2015	11 2	3995	1182	8	7%

The frequency of skills indication is slightly different when enterprises are divided into SMEs (including micro enterprises) and large enterprises. For SMEs, the first eight common skills indicated in offers are those listed above. The list was created when analyzing indications of all enterprises. For large enterprises, however, the following skills were less important: *foresee results of action undertaken* (9th in indications of large enterprises) and *update knowledge and develop professional skills* (10th in indications of large enterprises). From the perspective of large enterprises in the period analyzed, more important were the following skills: *be able to plan a working day* (5th in indications of large enterprises, whereas in case of small enterprises it was 22nd), *be able to organize own workstation* (6th in indications of large enterprises, whereas in case of small enterprises it was 23rd). As research in large enterprises co-creating the System for Professionals in the analyzed period shows, planning and organizational skills are of more importance. They are related to fulfilling the managing function in enterprises. These skills were not so significant for SMEs examined. The question may be posed – whether they are less important in this category of enterprises, or if persons preparing job offers undervalue functions of planning and organization in SMEs. Or perhaps, work positions for which these offers were published by large enterprises had a different specific character. These questions could be answered to some extent by analyzing the detailed content of the offers published in the system. Nevertheless, research results obtained by analyzing data from the system may be fodder for other research, pondering the reasons for discrepancies in the perception of key competences in enterprises of various sizes.

Apart from examining the importance of individual skills constituting key competences, the change in importance of these skills over time may also be examined in the System for Professionals. One may attempt at setting change trends of the skill importance by calculating the percentage of indications of a given skill in the total number of all indications. Figures 1-4 present importance changes over time of chosen common skills for professions.

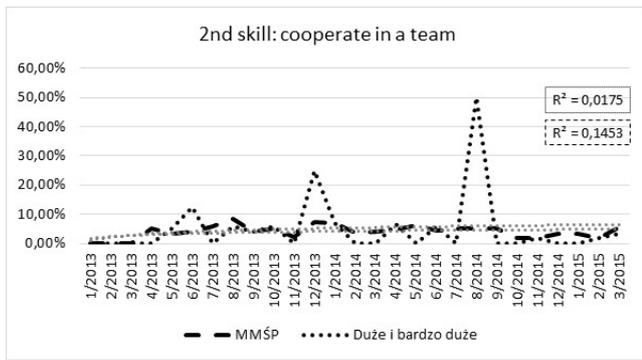


Fig. 1. Importance change over time for the skill *respect the rules of behaviour and ethics* in SMEs and large enterprises based on skill indications in offers of job, apprenticeship or training published in the System for Professionals.

Fig. 2. Importance change over time for the skill *cooperate in a team* in SMEs and large enterprises based on skill indications in offers of job, apprenticeship or training published in the System for Professionals.

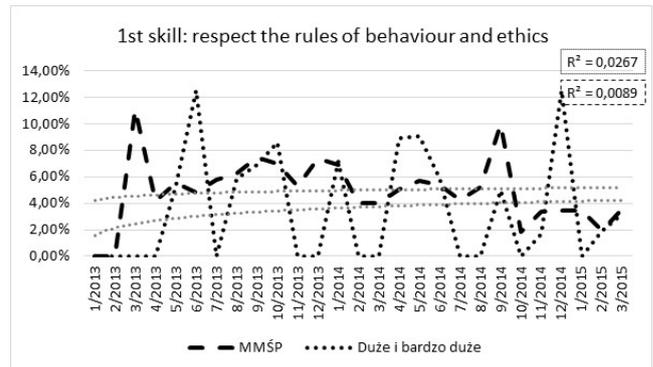


Fig. 4. Importance change over time for the skill *be creative and consistent in task realization* in SMEs and large enterprises based on skill indications in offers of job, apprenticeship or training published in the System for Professionals.

Trend lines on the graphs are solely examples. An in-depth analysis should decide whether the aspect of seasonal changes should be taken into consideration.

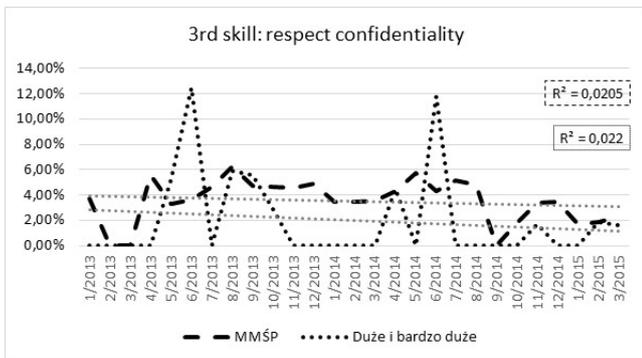
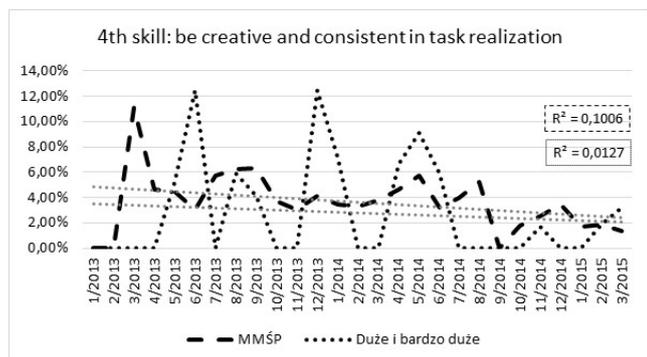


Fig. 3. Importance change over time for the skill *respect confidentiality* in SMEs and large enterprises based on skill indications in offers of job, apprenticeship or training published in the System for Professionals.



B. Space analysis

System for Professionals is dedicated to entrepreneurs looking for employees with certain competences. Among many expectations related to the detailed character of employee profile, key competences play a special role. Due to the varied availability of employees in the area of the Wielkopolska region, an important factor is the declaration of mobility by persons looking for a job. Frequently, the main factor when choosing a job is the distance between place of residence and place of work [26]. Mobility and related decisions of choosing a job very often influence the professional career of an employee. Employees' choices are reflected by employers' decisions connected with managing the business (opening a branch, remote work) and payroll policy (commuting and housing expenses compensation).

Authors of this article realized projects connected with the problem of mobility and optimization of information describing locations, which were used to create the System for Professionals [27], [28], [29]. In practical business management, the problem of availability of employees with certain key competences has a strong influence on human capital costs, and in consequence – on the future development of the business. More and more often, availability and cost of human potential expressed in qualifications possessed decide about the location of a

business. All these parameters are directly or indirectly reflected in the data gathered by System for Professionals. Both job offers and employees' job applications are identified by the address, allowing geolocation. It is essential due to costs of work of the candidate and the possibilities of building a steady bond with the enterprise. A declaration of the commuting distance is also an initial criterion of the choice of employees – optimizing the recruitment process. A varied structure of employment and flexible changes of competences in enterprises were taken into account while designing the System for Professionals. An employer may find forms of partial employment (including remote work), carry out internal recruitment in the Module of career counselling or increase employees' competences through E-learning. All these factors may be subjects of analyses in the System, especially of space analyses.

System for Professionals assumes not only day-to-day, detailed communication between employers and employees, but also the possibility to draw conclusions on the labour market. Based on the data from the System for Professionals, synthetic one-subject reports or detailed thematic studies are published.

An example of conclusions regarding the labour market in space grouping of the data from the System for Professionals may be the use of hierarchical clustering. Based on offers published by 918 employers (most of them published more than one offer), space analysis of competence needs in the region was carried out. Due to the limitations of this article, only potential analytical-prognostic possibilities of the System for Professionals are presented. For the needs of this article, competences for regions and sub-regions in Wielkopolska were aggregated. It is, however, possible, to analyze individual sub-regions.

Table 2. Space analysis of factors in sub-regions in search of similarities based on offers published in the System for Professionals – common skills for professions most frequently expected by employers.

Region	Area [km ²]	Unemployment [#of people]	Job offer	Common skills for professions
City Poznań	261	13 800	155	459
Leszno region	3 602	13 856	19	38
Kalisz region	5 786	24 823	150	224
Konin region	6 397	23 955	21	75
Piła region	6 459	21 995	29	140
Poznań region	9 541	41 018	207	282

Table 2 presents a chosen possibility to select homogenous groups of data based on Euclidean distances. Factors accepted for the analysis of sub-regions in the example concerned the unemployment rate in the sub-region, the number of job offers and the number of common skills for professions, most frequently expected by employers (the skills were discussed in point A). As a result of comparison of the qualities mentioned above, characterizing the labour market and the needs of employers' looking for employees through the System for Professionals, a diagram of average differences was created – fig. 5. Based on the analysis of concentration, the following typology groups can be defined: a) Kalisz region and Konin region, b) lower gravity between Leszno region and Piła region and c) no connections between the city of Poznań and Poznań region. The analysis carried out is a starting point for in-depth research, based on less synthetic

data, and for conclusions on the needs of business management.

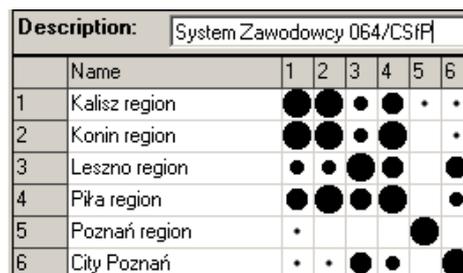


Fig. 5. An example of hierarchical clustering, grouping regions of similar typology based on offers of SMEs from the System for Professionals – common skills for professions most frequently expected by employers.

Table 3. Space analysis of factors in regions, searching for similarities, based on offers published in the System for Professionals – common skills for professions expected by employers.

Region	Area [km ²]	Common skills for professions	Unemployment [#of people x 1000]
Rawicz region	553.23	10	2.3
Śrem region	574.41	48	1.6
Jarocin region	587.7	138	3.2
Kępno region	600.39	149	1.1
Środa region	623.18	2	3.0
Grodzisk region	643.72	66	1.8
Wolsztyn region	680.03	14	1.4
Chodzież region	680.58	44	2.7
Września region	704.19	45	4.0
Pleszew region	711.91	27	2.6
Oborniki region	712.65	3	2.2
Krotoszyn region	714.23	174	3.1
Kościan region	722.53	8	2.1
Międzychód region	736.66	46	1.1
Ostrzeszów region	772.37	254	2.2
Leszno region	804.65	59	3.7
Gostyń region	810.34	4	3.6
Słupca region	837.91	119	3.7
Turek region	929.4	0	3.6
Koło region	1011.03	18	5.2
Nowytomysl region	1011.67	267	1.5
Wągrowiec region	1040.8	11	4.0
Szamotuły region	1119.55	276	2.8
Kalisz region	1160.02	99	5.8
Ostrów region	1160.65	206	5.4
Gniezno region	1254.34	124	6.6
Piła region	1267.1	117	5.8
Konin region	1578.71	106	12.9
City Poznan	1637.81	1675	4.0
Złotów region	1660.91	3	4.1
Czar-Trzci region	1808.19	100	4.0
Poznań region	1899.61	681	6.0

Table 3 presents parameters characteristics of the labour market. Among the qualities mentioned, common skills for professions expected by employers and the unemployment rate are given.

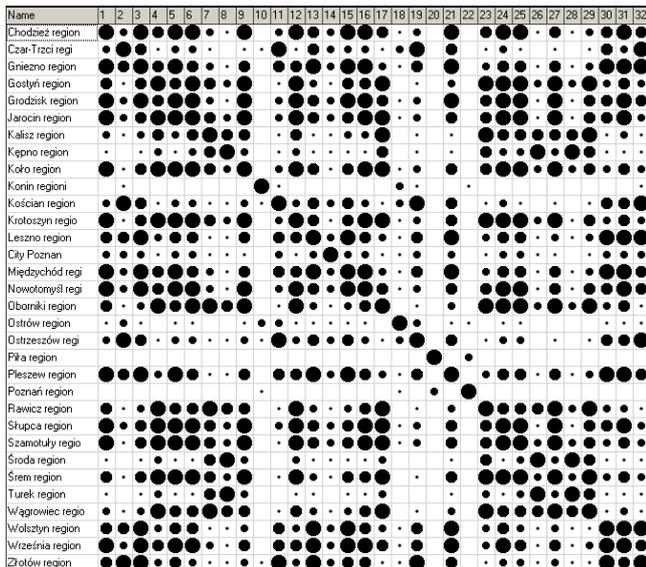


Fig. 6. An example of hierarchical clustering, grouping regions of similar typology based on System for Professionals – common skills for professions expected by employers.

Based on groups of sub-regions according to chosen factors, fig. 6 shows several segments showing similarities. Three groups possess qualities enabling mutual gravity and four regions show individuality. This type of grouping may be an initial selection for further analysis. Another step should be to find special factors for similar groups of regions. Analyses of this type may help entrepreneurs decide on the location of the branch or explain difficulties connected with finding a given competence in that region. Local authorities may use such analyses to forecast the educational potential or the necessity to introduce courses and trainings developing competences. As the System develops and data grows, entrepreneurs and local authorities will be offered more precise analyses and forecasts.

V. CONCLUSION

The problem of systems of computer aided for business environment monitoring is present in the literature for a long time [30]. Development of Information Technology and Management is responsible for their improvement. The article presented only chosen possibilities of how to use the System for Professionals. At this stage of development, analyses may be enhanced with the factor expressing the expected level of certain skills. It is also possible to compare evaluation of employers and candidates. One of the most important activities is currently improvement of quality of data used in research. In the nearest future, it is planned to:

- 1) increase the number of users to enable conclusions for sub-regions, poviats (counties) or big cities,
- 2) transfer the system to other regions and target groups of users in order to lower unit costs of the system and also to examine the influence of cultural differences as well as demographic, social, geographical and other characteristics on the evaluation of the importance of key competences,
- 3) attribute skills from the system to key competences in a less ambiguous way, to create aggregated factors,
- 4) prepare automatic reports and develop a system of fast information about changes in the importance of a skill over time and space,

- 5) further popularize the system among entrepreneurs,
- 6) prolong the lifespan of users in the system through the development of mechanisms of building relationships with them.

On 16 April 2015, the Board of the Wielkopolskie Voivodeship invited Poznań University of Technology to a seven-year partnership, which will involve, among others, development of the System for Professionals, including development of the research system. Other activities are undertaken to encompass groups of candidates other than students and graduates of secondary vocational schools.

REFERENCES

- [1] M. Szafrński, *Zarządzanie akceleracją tworzenia zasobów wiedzy w przedsiębiorstwach*, Wydawnictwo Politechniki Poznańskiej, Poznań 2015, pp. 163-167.
- [2] E. Chmielecka, Z. Marciniak, A. Kraśniewski, *Krajowe Ramy Kwalifikacji dla polskiego szkolnictwa wyższego*, w: *Autonomia programowa uczelni. Ramy kwalifikacji dla szkolnictwa wyższego*, red. E. Chmielecka, Ministerstwo Nauki i Szkolnictwa Wyższego, Warszawa 2010.
- [3] *From Berlin to Bergen*, General Report of the Bologna Follow-up Group to the Conference of European Ministers Responsible for Higher Education (2005), Norwegian Ministry of Education and Research, Bergen, 19-20 May.
- [4] Ch. Evans, *Zarządzanie wiedzą*, PWE, Warszawa 2005.
- [5] *The Knowledge-based Economy*, Organisation for Economic Co-operation and Development, Paris 1996.
- [6] G. Probst, S. Raub, K. Romhardt, *Zarządzanie wiedzą w organizacji*, Oficyna Ekonomiczna, Kraków 2002.
- [7] *Glosariusz terminów i pojęć używanych w europejskich programach współpracy w dziedzinie edukacji*, opr. E. Kolanowska, Fundacja Rozwoju Systemu Edukacji, Warszawa 2010.
- [8] Rozporządzenie Ministra Edukacji Narodowej z dnia 7 lutego 2012 w sprawie podstawy programowej kształcenia w zawodach, Dz. U., 2012, poz. 184 (przepisy polskie).
- [9] Zalecenie Parlamentu Europejskiego i Rady nr 2006/962/WE z dnia 18 grudnia 2006 r. w sprawie kompetencji kluczowych w procesie uczenia się przez całe życie (Dz.U. L 394 z 30.12.2006).
- [10] *Edukacja oparta na kompetencjach dla szkół z całej Europy*, (2015 April 16) Available: http://www.schooleducationgateway.eu/pl/pub/newsevents/competence-based_education.htm.
- [11] *Developing Key Competences at School in Europe: Challenges and Opportunities for Policy* Eurydice Report, European Commission/EACEA/Eurydice, 2012. *Developing Key Competences at School in Europe: Challenges and Opportunities for Policy*. Eurydice Report. Luxembourg: Publications Office of the European Union.
- [12] M. Szafrński, K. Grupka, M. Goliński, *Program akceleracji wiedzy technicznej i matematyczno-przyrodniczej w Polsce*, Wyd. PP, Poznań, 2008.
- [13] S. Truskolaski, *Znaczenie transferu wiedzy w działalności innowacyjnej przedsiębiorstw*, Wydawnictwo Difin, Warszawa 2014.
- [14] A. La Rocca, I. Snehota, *Relating in business networks: Innovation in practice*, *Industrial Marketing Management*, 43, 2014 pp. 441–447.
- [15] P. Sok, A. O’Cass, K. Mory Sok *Achieving superior SME performance: Overarching role of marketing, innovation, and learning capabilities*, *Australasian Marketing Journal*, 21, 2013 pp. 161–167.
- [16] K. Sacha, *Inżynieria oprogramowania*, Wydawnictwo Naukowe PWN SA, Warszawa 2010, p. 334.
- [17] P. Berłowski, *Wskaźniki zarządzania kapitałem ludzkim w Elektrobudowie SA*, *Personel i zarządzanie*, Wyd. Grupa Infor pl, nr 2 2015, pp 70-73.
- [18] G. Filipowicz, *Zarządzanie kompetencjami: perspektywa firmowa i osobista*, Wolters Kluwer, Warszawa, 2014.
- [19] M. Spychała, *Analysis and Improvement of the Process Engineer’s Levels of Competence in a Manufacturing Company*, *Logistics Operations, Supply Chain Management and Sustainability*, ed. Paulina Golińska, Springer International Publishing, 2014, pp 395-409.

- [20] M. Armstrong, S. Taylor. *Armstrong's handbook of human resource management practice*, London; Philadelphia Kogan Page, 2014, pp 573-582
- [21] M. Leśniewski, *Kapitał intelektualny w kształtowaniu zrównoważonego rozwoju przedsiębiorstw*, Ekonomika i organizacja przedsiębiorstw, Wyd. IOiZwP Orgmasz, 2015, pp. 14-25.
- [22] A. Różański, M. Bajor, B. Kozak, *Gotowość prorozwojowa osób świadczących prace w oparciu o elastyczne formy zatrudnienia*, Przegląd organizacji, 11, 2014, pp. 22-28.
- [23] M. Goliński, K. Grupka, M. Szafrąński, *Akcelerator Wiedzy Technicznej@ - projektowanie przyszłości*, Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej Nr 37, I Konferencja e-Technologies in Engineering Education eTEE'2014, Politechnika Gdańska, Gdańsk 2014, pp 81-84.
- [24] Uchwała nr 1979/08 Zarządu Województwa Wielkopolskiego z dnia 20 listopada 2008 roku w sprawie przyjęcia „Koncepcji organizacyjnej kształcenia kadr kwalifikowanych i kształcenia ustawicznego w Wielkopolsce na poziomie zasadniczej szkoły zawodowej, technikum, szkoły policealnej i kolegium, dokształcania, doskonalenia i doradztwa”, stanowiącej załącznik do uchwały.
- [25] M. Szafrąński, *Praktyki zawodowe – narzędzie zarządzania wiedzą wspomagające obniżanie kosztów w przedsiębiorstwach*, Przegląd organizacji, No. 1/2015, pp. 29-35.
- [26] M.A. Carree, K. Kronenberg, *Locational Choices and the Costs of Distance: Empirical Evidence for Dutch Graduates*, Spatial Economic Analysis, Journal of the Regional Studies Association and the RSAI British and Irish Section, Vol 9 no 4. 2015.
- [27] M. Goliński, *The use of web application Mobilne miasto [Mobile city] in the conveyance of information about urban space in the system human factor – technology*, Advances in social and organizational factors / ed. by Peter Vink : AHFE Conference, Advances in Human Factors and Ergonomics, 2014 pp 206-216.
- [28] *Integrated support system for access to information in urban space with use of GPS and GIS systems* M. Goliński, M. Szafrąński (ed.), , Wyd. Politechniki Poznańskiej, Poznań, 2012.
- [29] M. Goliński, M. Szafrąński, M. Graczyk, W. Prussak, T. Skawiński, : Technological and organizational determinants of information management in the urban space (based on scientific research), ACM ICUIMC 2012, February 20–22, Kuala Lumpur, Malaysia.
- [30] D. Jelonek, *Systemy komputerowego wspomaganie monitorowania otoczenia przedsiębiorstwa*, Wyd. WZ P.Cz., Częstochowa 2002.

Data assimilation method coupled with the numerical simulation of the ocean dynamics

Konstantin P. Belyaev, Andrey A. Kuleshov, Clemente A. S. Tanajura, Natalia P. Tuchkova

Abstract—A new data assimilation method based on the properties of stochastic diffusion processes is derived. The method combines variational and statistical approaches commonly used in data assimilation theory. The proposed scheme minimizes the variance of the trajectory of a diffusion process considered as the limit process of sequence of applications of data assimilation technique in conjunction with the numerical model. This scheme differs from the Kalman-filter and appears as the consequence of the path-of-least-resistance principle, reducing the assimilation problem to a system of linear equations in model phase-space. This method is applied into the HYbrid Coordinate Ocean Model (HYCOM) and assimilates satellite sea level anomaly data from the Archiving, Validating and Interpolating Satellite Ocean Data (AVISO) over the Atlantic Ocean. This method is applied to correct the model dynamics in Atlantic. Several numerical experiments have been performed. The experiments show that the method substantially changes the synoptic and mesoscale structure of ocean dynamics.

Keywords—Diffusion stochastic processes, ocean data assimilation, Kalman gain matrix, sea level anomaly, ocean dynamics, synoptic and mesoscale structure.

I. INTRODUCTION

Data assimilation (DA) techniques are common in operational oceanography and weather and climate forecasting [1]. The goal of those techniques is to combine the ocean and/or atmosphere background state produced by a numerical model with independent observations and to create a new state, the so-called objective analysis. This analysis should represent the ocean and/or atmosphere physics better than the pure model background. It should provide analyzed variables closer to the observations in any reasonable sense and, therefore, should produce a better forecast if used as the numerical model initial condition. The skill and usefulness of any assimilation technique can be assessed by the deviation of the analysis with respect to independent data and also by the computational cost and programming, among other aspects.

Theoretical part of this work (ch. II) is performed with financial support of Russian Science Foundation grant 14-11-00434.

K. P. Belyaev, Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow, Russia (e-mail: kbel55@yahoo.com).

A. A. Kuleshov, Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Moscow, Russia (e-mail: andrew_kuleshov@mail.ru).

C. A. S. Tanajura, Oceanographic Modeling and Observation Network (REMO), Center for Research in Geophysics and Geology, Federal University of Bahia (UFBA), Salvador, Brazil (e-mail: clemente.tanajura@gmail.com).

N. P. Tuchkova, Dorodnitsyn Computational Center, Russian Academy of Sciences, Moscow, Russia (e-mail: natalia_tuchkova@mail.ru).

Most of DA approaches are divided into two large groups, namely the variational or functional schemes and statistic or dynamical-statistic schemes. The first group is represented by the three- or the four-dimensional variational schemes (3D-Var or 4D-Var) [2]-[4], while the statistic approach is mostly represented by the Kalman-filter scheme and its simplified version the Ensemble Optimal Interpolation scheme (EnOI) [5], [6]. There are several original hybrid schemes combining aspects of both approaches [7], [8].

The main difference of these two main ideologies is on how to minimize the model error, i.e., the difference between the model and unknown truth after assimilation. The 4D_Var approach seeks the analysis as the optimal initial condition and turns the model trajectory starting from this initial condition as close as possible to observations. On the other hand the statistical approach seeks the optimal Kalman gain (weight matrix) to keep the balance between the model solution and observations after DA. Therefore, 4D-Var changes the initial condition at the beginning of the observational window while the Kalman filter changes the model state instantaneously and independently of the initial condition that produced this model state. Because of this reason the Kalman filter realization is in general simpler than the 4D-Var. However, the Kalman filter does not consider the assimilation as a process, since once the model error is estimated the correction of the model state is made at once. Alternatively, the 4D-Var considers the time evolution of the model error in the observational window.

The current work derives and uses another DA scheme, a hybrid scheme, in which the Kalman gain is sought with explicit dependence on the initial condition that produced the model background state and the difference between model and observations at the assimilation time. Unlike the Kalman filter scheme, this approach uses the general path-of-least-resistance principle, which in special cases, is reduced to the minimum variance problem in conjunction with a known tendency shown by the observational data. It should be mentioned that this tendency may not coincide with the model tendency. In the present work, the observational tendency is calculated in phase-space and details are presented below. Theoretically the proposed scheme is based on the properties of diffusion stochastic processes but for its realization it is sufficient to know the model gradient and error covariance matrix as in both the 4D-Var and the Kalman-filter.

The numerical ocean modeling is an important part in geophysics. Starting from the fundamental works [9], [10] this scientific area passes a great way and now it is one of the

most important and developed direction in oceanology and Earth studies. DA experiments were here performed with the HYbrid Coordinate Ocean Model (HYCOM) [11], [12] version 2.2.14 configured over the Atlantic Ocean from 79°S until 55°N and 100°W until 20°E, including Caribbean sea but excluding the Mediterranean Sea and the Pacific Ocean. Sea level anomaly data from the Archiving, Validating and Interpolating Ocean data (AVISO) [available from www.aviso.altimetry.fr] were utilized. In this direction the current work continues the studies previously published in [13], [14].

Since this paper mostly focuses on geophysical studies all necessary mathematical formalism is presented very specifically for the geophysical community. Abstract formulations and redundant generalizations were avoided. The present work is a fragment of a general data assimilation system under development for operational purposes within the Brazilian Oceanographic Modeling and Observation Network (REMO) [www.rederemo.org].

Therefore, the main goals of the current study are: (i) to derive and develop a new hybrid data assimilation method; (ii) to show the feasibility of its application; (iii) to investigate the impact of the assimilation of sea level anomalies with the proposed method and (iv) to compare its performance with the EnOI technique.

II. MATHEMATICAL FORMULATION

Let the ocean numerical model be integrated on the time interval $[0, T]$ on the gridded domain and X be the model state vector that encompasses for instance, temperature, salinity, sea surface height (SSH) and others. Let the number of grid points be denoted as N_g and the number of model variables be N_{mv} . Then the dimension of X will be $N_g \times N_{mv} = r$. Let Y be the observational vector and N_o the number of observations each with N_{ov} variables, for instance, temperature and sea surface height anomaly (SSHA). Then the total dimension of the observational vector state is $N_o \times N_{ov} = N$. Normally, $N_{mv} \geq N_{ov}$ since not all model variables are observed. As it is usual in DA theory two ocean state vectors are introduced, namely the analysis and background states, X_a, X_b , respectively. They are linked by the equality

$$X_a = X_b + K_{i+1}(Y_{i+1} - HX_b), \quad (1)$$

where matrix K is the so-called Kalman gain with dimension $r \times N$ while matrix H with dimension $N \times r$ is the observation operator that projects the model space into the observational space.

The interval $[0, T]$ is broken down into subintervals $[t_i, t_{i+1}]$, $0 = t_0 < t_1 < \dots < t_l = T$, and at times $t_i, i = 1, \dots, l$ assimilation is executed according to formula (1), so at all moments t_i one has the equality

$$X_{a,i+1} = X_{b,i+1} + K_{i+1}(Y_{i+1} - HX_{b,i+1}). \quad (2)$$

Let the model background or first-guess be given on each of those subintervals be denoted as $X_{b,i+1} = F(X_{a,i})$.

Therefore

$$X_{a,i+1} = F(X_{a,i}) + K_{i+1}(Y_{i+1} - HF(X_{a,i})). \quad (3)$$

Finally, if the model $F(x)$ is represented as

$$F(x) = x + \int_{t_i}^{t_{i+1}} \Lambda(x_i, \tau) d\tau \text{ for some } \Lambda, \text{ Eq. (3) is turned}$$

into

$$X_{a,i+1} = X_{a,i} + \int_{t_i}^{t_{i+1}} \Lambda(X_{a,i}, \tau) d\tau + K_{i+1} \left(Y_{i+1} - HX_{a,i} - H \int_{t_i}^{t_{i+1}} \Lambda(X_{a,i}, \tau) d\tau \right). \quad (4)$$

Since Eq (4) is written in term of analyses, hereafter the subscript a will be omitted.

In [7] it was shown that under certain physically reasonable conditions the process $X_{a,i}, i = 1, \dots, l$ can be approximated by the stochastic diffusion process of the following type

$$dX(t) = (I + KH)\Lambda dt + (KQK')dW, \quad (5)$$

where X is the identity matrix, $Q = E(Y - HX)(Y - HX)' + R$ is the model error covariance matrix plus observational error covariance matrix. The latter is supposed to be independent of the model, dW is the standard notation of a white noise random variable and the apostrophe $'$ denotes the transpose of a vector or a matrix. Time t belongs to the subinterval $t_i < t < t_{i+1}$. Furthermore the subscript i will be omitted if it does not lead to any uncertainty. As usual the model is supposed to be unbiased with respect to observations, i.e., $E(Y - HX) = 0$ where symbol E stands for the ensemble average or mathematical expectation.

As it is seen in (5) this process is determined for all grid points and all variables. Without loss of generality it is possible to suggest that matrices KH and Q are invertible, so there exist $(KH)^{-1}$ and Q^{-1} . This condition simply means that each model variable and observational variable are linearly independent, i.e., no variable can be linearly represented by the others.

Now the assimilation problem may be formulated as follows: find out the Kalman gain K by minimizing in the sense of matrix norm the diffusion matrix KQK' under the

conditions that the drift vector or tendency $(I + KH)\Lambda$ is given and equal to C . According to this formulation vector C is an r -dimensional vector given at all grid points and having certain value for each model variable. In order to mathematically solve this optimization problem, the theory of conditional extremes will be applied [15]. According to this theory a functional L is constrained as follows:

$$L(K, \varphi) = KQK' + [(I + KH)\Lambda - C]\varphi, \quad (6)$$

where φ is an auxiliary unknown r -dimensional vector string, the so-called Lagrange multiplier vector. The minimization of this functional leads to equations

$$KQ + \frac{1}{2}\varphi'(H\Lambda)' = 0, \quad (7)$$

$$(I + KH)\Lambda = C.$$

Equations (7) are matrix equations which are equivalent to $(r+1) \times N$ linear scalar equations containing the same number of unknown variables. Since matrices KH and Q are invertible this system of linear equations always has a unique solution.

The solution of system (8) can be obtained explicitly. It is the following

$$K = \frac{(C - \Lambda)(H\Lambda)'Q^{-1}}{(H\Lambda)'Q^{-1}H\Lambda}. \quad (8)$$

The great advantage of the presented theory is the opportunity to calculate not only the analysis but its distribution as well. This distribution will be given by Fokker-Planck equation

$$\frac{\partial p(t, x)}{\partial t} = -\frac{\partial(I + KH)}{\partial x} + \frac{1}{2} \frac{\partial^2(KQK')}{\partial x^2}, \quad (9)$$

where $p(t, x)$ is the probability density. This equation is solved under initial condition $p(t_0, x) = p_0(x)$ and boundary condition $p(t, \pm\infty) = 0$, where $p_0(x)$ is an a priori given function.

III. RESULTS OF NUMERICAL SIMULATION

This method has been applied in conjunction with the ocean circulation model HYCOM [11], [12] with its recent configuration, version 2.2.14 [13], [14]. In the current version the model grid is configured over Atlantic from Antarctic up to 55°N, it has a spatial resolution 0.25° in eastward-westward direction and varied spatial resolution in southern-northern direction with minimum distance 0.25° between 10°S and 10°N. In vertical this version has 21 density layers as reference coordinate, total 480×760×21 grid points. Model calculates

109 independent variables in each grid points, namely sea surface heights, 3 barotropic variables (horizontal velocities and barotropic pressure) and 5 baroclinic variables for each reference level, namely horizontal velocities, layer thickness temperature and salinity. As entered information only sea surface height anomalies are taken from archive AVISO, observed and daily recorded sea surface heights minus their temporal average over 8 years, 2002-2009.

Several preprocessed steps have been performed before assimilation experiments. Since model is assumed to be theoretically unbiased the observed bias must be removed from observations. This is done according to the so-called along-track strategy bias removing [14]. Model data and observed data independently over the each observational satellite track are averaged and then their difference (observational averaged minus model average over each track) is subtracted from observations. Also, to minimize the number of observations needed to assimilate all data lying beyond the considered domain as well as all data occurred on islands or at the continental model cell are through off.

The observational trend C and matrix Q also are constrained on the preprocessed step. For each grid point the circle with radius 0.25 grid distance is taken and all observations occurred within this circle are averaged and assign to this grid point. Then the value C at this grid point will be equaled to new constrained average minus model value on the previous time step at this point. The procedure to constrain the matrix Q is more complicated. It is defined through the anomaly strategy as it is usually done in Ensemble OI method. Previously model is forced by NCEP climate reanalysis for 40 years. Then last 10 years of this run is taken as the basis. Over this 10 year the average value and anomalies (real model output minus average) are recorded for each day. Then the covariance Q is

calculated as $Q = \frac{1}{N_{ens}} \sum_{i=1}^{N_{ens}} (\theta_i - x)(\theta_i - x)'$ where θ_i is

anomaly at each observational point and $i=1, \dots, N_{ens}$, x is the model value interpolated onto this point on previous time-step, $N_{ens}=50$. To assimilate data at specific date, 5 nearby days are taken equidistantly, for instance, if assimilation is doing at April 7 the anomalies are taken at April 1,4,7,10, 13. As it is followed from this construction the matrix Q is $N \times N$ symmetric matrix and, as it was said before, it is invertible.

Once these preprocessed parameters are defined the assimilation experiments with the aforementioned scheme have been carried. Experiments started at April 1, 2011 and last 15 days with daily assimilation. Matrix K is constrained with respect to (9) and analysis $X(t+dt)$ is found out according to (4). This is done for all x .

Experiments show that the applied DA method works properly and really assimilates data. Figure 1(a-c) demonstrates respectively SSHA field after assimilation (analysis), before assimilation (background) and their difference on the last day of experiment (15-th day). One may see in Fig. 1(a) the mesoscale structure of SSHA clearly

pronounced in Northern part of domain (Gulfstream zone) and also in Drake Passage zone and Brazil-Malvinas zone in Southern Atlantic. Amplitudes of anomalies reach up to 0.6 which seems as a very large impact in synoptic scale. Fig 1(b) reproduces the same structure as Fig.1(a) but with smaller amplitudes and weakly pronounced mesoscale dynamics.

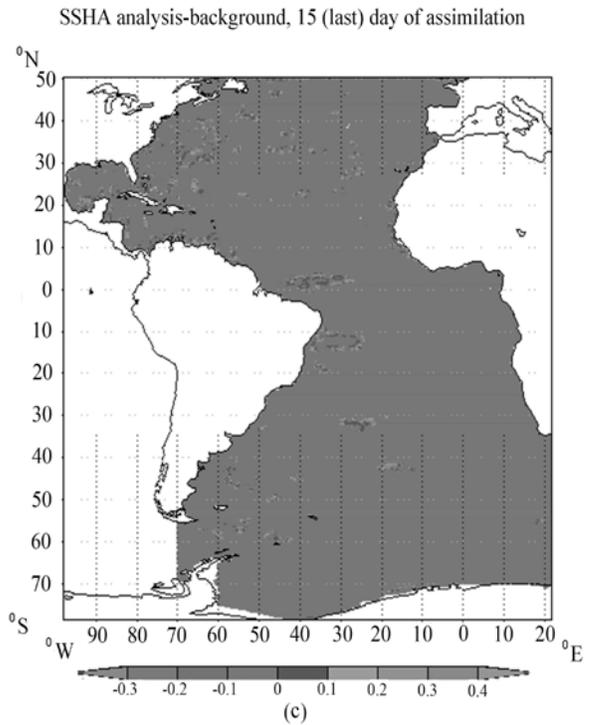
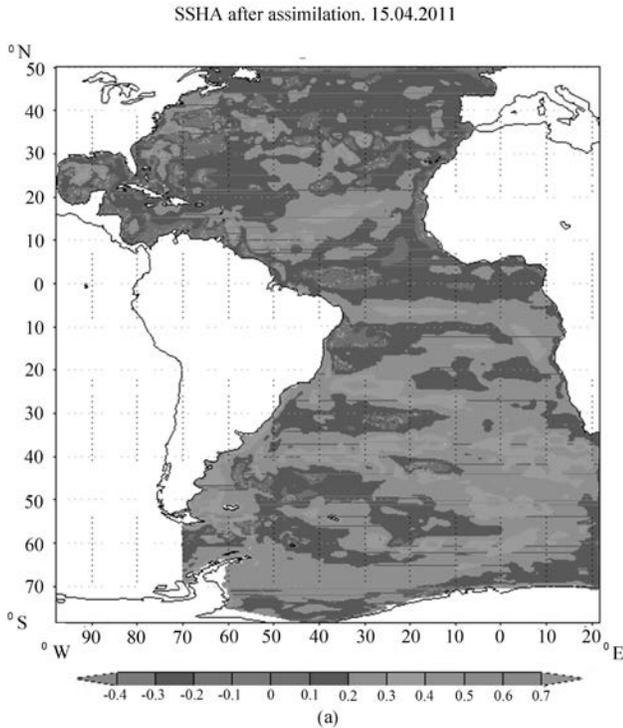


Fig. 1 SSHA fields on 15-th day of assimilation (a) after assimilation (b) before assimilation (c) their difference.

It is seen very well in Fig.1(c) (analysis-background). There are very well pronounced eddies with the amplitude up to 0.3 in Northern and Southern parts of Atlantic both with positive and negative dynamics. As a conclusion one may asserts that the DA captures and even aggregates the mesoscale and synoptic structure of SSHA rather than the control calculations. The exceptions are the eddies in Gulfstream zone which are simulated in both experiments.

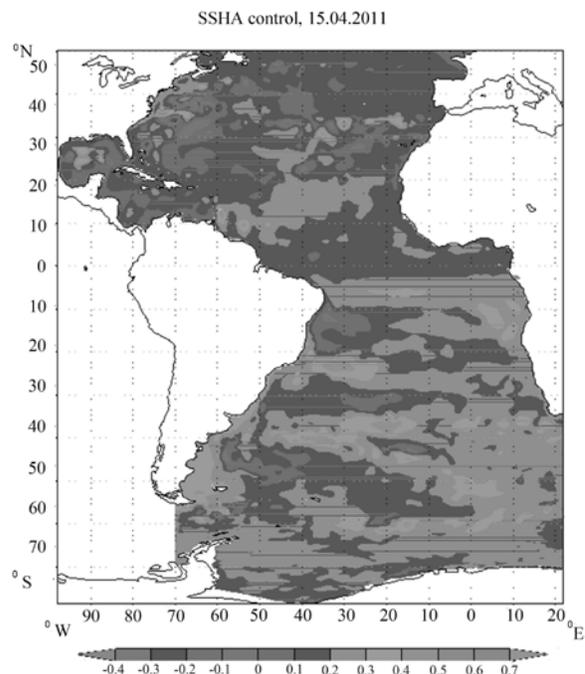
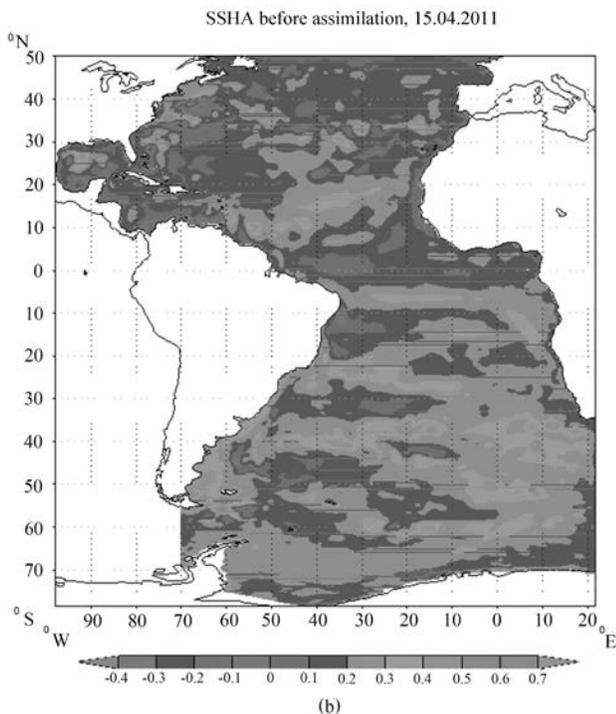


Fig. 2 SSHA control on 15 th day of assimilation.

Figure 2 shows the map of SSHA in twin experiment (control) on the same day.

IV. CONCLUSION

The new DA method is theoretically derived and practically realized in the current work. This method is quite simple in its applications and allows capturing the real synoptic and mesoscale oceanic structure. This work shows that the DA of altimetry reduces the model error and approaches the model trajectory close to the observations. Also, this study shows that it is possible to constrain the error bounds for the resulted trajectories.

Only SSHA data are assimilated in this work, however it is possible to expand this method to all observed (measured) data.

REFERENCES

- [1] M. Ghil, P. Malnotte-Rizzoli, "Data assimilation in meteorology and oceanography", *Adv. Geophys.*, vol. 33, 1991, pp. 141–266.
- [2] S. Cohn, "An introduction to estimation theory", *J. Meteor. Soc. Japan*, vol. 75, ch. 1B, 1997, pp. 257–288.
- [3] V. I. Agoshkov, V. M. Ipatova, V. B. Zalesnyi, E. I. Parmuzin, V. P. Shutyaev, "Problems of variational assimilation of observational data for ocean general circulation models and methods for their solution", *Izv. Atmos. Ocean. Phys.*, vol. 46, no. 6, 2010, pp. 677–712.
- [4] V. B. Zalesny, G. I. Marchuk, "Modeling of the World Ocean circulation with the four-dimensional assimilation of temperature and salinity fields", *Izv. Atmos. Ocean. Phys.*, vol. 48, no. 1, 2012, pp. 15–29.
- [5] G. Evensen, "The ensemble Kalman filter: Theoretical formulation and practical implementation", *Ocean Dyn.*, vol. 53, 2003, pp. 343–367.
- [6] J. Xie, J. Zhu, "Ensemble optimal interpolation schemes for assimilating Argo profiles into a hybrid coordinate ocean model", *Ocean Modelling*, vol. 33, 2010, pp. 283–298.
- [7] C. A. S. Tanajura, K. Belyaev, "A sequential data assimilation method based on the properties of diffusion-type process", *Appl. Math. Model.*, vol. 33, 2009, pp. 2165–2174.
- [8] C. Lorenc, N. E. Bowler, A. M. Clayton, S. R. Pring, D. Fairbairn, "Comparison of Hybrid-4DVar and Hybrid-4DVar data assimilation methods for Global NW", *Mon. Wea. Rev.*, vol. 143, 2015, pp. 212–229.
- [9] K. Bryan, "A numerical method for the study of the circulation of the world ocean", *J. of Comp. Phys.*, vol. 4, 1969, pp. 347–376.
- [10] S. Sarkisyan, *Theory and calculation of ocean currents*. Leningrad: Gidrometeoizdat, 1966 (Russian).
- [11] R. Bleck, D. B. Boudra, "Initial testing of a numerical ocean circulation model using a hybrid quasiisopycnal vertical coordinate", *J. Phys. Oceanogr.*, vol. 11, 1981, pp. 755–770.
- [12] R. Bleck, "An oceanic general circulation model framed in hybrid isopycnic-Cartesian coordinates", *Ocean Model.*, no. 4, 2002, pp. 55–88.
- [13] K. Belyaev, C. A. S. Tanajura, N. Tuchkova, "Comparison of Argo drifter data assimilation methods for hydrodynamic models", *Oceanology*, vol. 52, no. 5, 2012, pp. 523–615.
- [14] C. A. S. Tanajura, L. N. Lima, "Assimilation of sea surface height anomalies into HYCOM with an optimal interpolation scheme over the Atlantic ocean Metarea V", *J. Bras. Geofis.*, vol. 31, 2013, pp. 257–270.
- [15] F. P. Vasil'ev, *Metody optimizatsii*, vol. 1. Moscow, 2011 (Russian).

Implementation of a kinetically-based algorithm for porous medium flow simulation on hybrid supercomputers

Andrew A. Kuleshov, Natalia G. Churbanova, Anastasiya A. Lyupa, and Marina A. Trapeznikova

Abstract—The work deals with further development of an original approach to porous media flow simulation using a kinetically-based model and explicit methods for its numerical implementation. Flow of three-phase slightly compressible fluid is under consideration. The computational algorithm is adapted to hybrid supercomputers with graphics accelerators and demonstrates high parallelization efficiency on test problems concerning infiltration processes.

Keywords—Multiphase porous medium flow, quasigasdynamic system of equation, automatic data partitioning, parallel implementation.

I. INTRODUCTION

MATHEMATICAL modeling of multiphase fluid flows in porous media is necessary for solving many practically important problems, in particular, investigations of processes of hydrocarbon recovery and processes of contaminant infiltration into the soil. It is well known that numerical simulation of these large-scale processes is very time-consuming and impossible without the employment of high-performance computer systems. Nowadays the rapid growth in the computer performance is mainly achieved due to the use of hybrid architectures including multicore CPUs and different accelerators like graphics processing units (GPU). Usage of GPU for general-purpose computations is a perspective modern trend to solve problems of mathematical physics with high accuracy for the reasonable time. However such architectures cause serious difficulties in the software development. Computational algorithms with logical simplicity, for example, explicit finite-difference schemes can

Theoretical part of this work (ch. II) is performed with financial support of Russian Science Foundation (RSF), grant 14-11-00549. Parallel implementation and test predictions (ch. III) are done with financial support of Russian Foundation for Basic Research (RFBR), grants 15-01-03445 and 15-01-03654.

A. A. Kuleshov, Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Moscow, Russia (e-mail: andrew_kuleshov@mail.ru).

N. G. Churbanova, Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Moscow, Russia (e-mail: nataimamod@mail.ru)

M. A. Trapeznikova, Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Moscow, Russia (e-mail: mtrapez@yandex.ru).

A. A. Lyupa, Moscow Institute of Physics and Technology (State University), Moscow, Russia (e-mail: nastenka.aesc@gmail.com).

be adapted easily to hybrid supercomputers and allow to exploit them more efficiently.

In the present paper a new approach to porous medium flow simulation is discussed: the model constructed by the analogy with the quasigasdynamic system of equations [1], [2] is applied to the case of three-phase fluids and assumes implementation by explicit difference schemes with rather a mild stability condition. The parallel software library for modeling processes in the subsurface [3] is supplemented now with new computational modules including an original procedure of automatic data partitioning among processing units to provide the optimal load balancing.

The developed tools are verified by solving a number of 3D test problems of three-phase infiltration over computational grids up to 40 million points. High parallelization efficiency is achieved on a classical cluster as well as on a GPU-based cluster.

II. GOVERNING MODEL AND COMPUTATIONAL ALGORITHM

The mathematical model of multiphase porous media flows is developed by the analogy with the kinetically-consistent finite difference (KCFD) schemes and the related quasigasdynamic (QGD) system of equations [1]. We start from the classical model of slightly compressible fluid flow in a porous medium [4] written as follows:

$$\frac{\partial \rho}{\partial t} + \operatorname{div} \rho \mathbf{u} = 0, \quad (1)$$

$$\mathbf{u} = -\frac{K}{\mu} \operatorname{grad} P, \quad (2)$$

$$P = P_0 + \beta(\rho - \rho_0). \quad (3)$$

Here ρ is the density, P is the pressure, \mathbf{u} is the Darcy velocity, K is the absolute permeability, μ is the dynamic viscosity, β is the compressibility factor, P_0 and ρ_0 are constant base values of the pressure and the density.

On the basis of the so-called principal of minimal sizes [1] we assume the existence of some minimal reference space and time scales which act as lower limits of description details. For example, in gas dynamics the free path of a molecule and the

time interval between molecule collisions are values on such a scale. For porous media flows the minimal reference length l is a distance at which the rock microstructure is negligible (this is a distance of the order of hundred rock grain sizes), the minimal reference time τ is time for inner equilibrium establishing in the volume with the reference size l . Microscopic equations for filtering fluids are averaged over these reference scales [2]. Ideas applied at the derivation of the QGD system are combined now with the differential approximation technique. Consequently the traditional parabolic continuity equation (1) gets a regularizing term and a second order time derivative with small parameters having the sense of the corresponding minimal reference scales:

$$\frac{\partial \rho}{\partial t} + \tau \frac{\partial^2 \rho}{\partial t^2} + \operatorname{div} \rho \mathbf{u} = \operatorname{div} \frac{lc}{2} \operatorname{grad} \rho, \quad (4)$$

where c is the magnitude of the order of the sound speed in fluid. The additional diffusion term in the right-hand side guarantees solution smoothing on length l .

Hyperbolic equation (4) can be approximated by the conditionally stable three-level explicit scheme of the second order of approximation on time and on space using central differences for discretization of the convective term "div $\rho \mathbf{u}$ ".

In previous works, e.g. [2], it was shown that instead of the strong time-step restriction of explicit schemes for parabolic equations

$$\Delta t \sim \Delta h^2 \quad (5)$$

the scheme for (4) has the next milder stability condition:

$$\Delta t \sim \Delta h^{3/2}, \quad (6)$$

where Δh is the spatial grid step.

In paper [2] the kinetically-based model for two-phase fluid flow in a porous medium was proposed. The given below system of equations introduces the generalization to the case of three-phase fluid flow (the subscript α indicates the phase – water (w), Non-Aqueous Phase Liquid (n) or gas (g)):

$$\begin{aligned} \varphi \frac{\partial(\rho_\alpha S_\alpha)}{\partial t} + \tau \frac{\partial^2(\rho_\alpha S_\alpha)}{\partial t^2} + \operatorname{div} \rho_\alpha \mathbf{u}_\alpha = \\ = q_\alpha + \operatorname{div} \frac{l_\alpha c_\alpha}{2} \operatorname{grad}(\rho_\alpha S_\alpha), \end{aligned} \quad (7)$$

$$\mathbf{u}_\alpha = -K \frac{k_\alpha}{\mu_\alpha} (\operatorname{grad} P_\alpha - \rho_\alpha \mathbf{g}), \quad (8)$$

$$\rho_g = \rho_{0g} \frac{P}{P_{0g}}, \quad \rho_\alpha = \rho_{0\alpha} [1 + \beta_\alpha (P_\alpha - P_{0\alpha})], \quad \alpha = w, n, \quad (9)$$

$$\sum_\alpha S_\alpha = 1. \quad (10)$$

Here φ is the porosity, S_α is the α -phase saturation, q_α is the source of fluid, k_α is the relative phase permeability, \mathbf{g} is the gravity vector.

The system is completed by relationships for the relative phase permeability. In the case of three phases k_α is determined by Stone approach after K. Aziz and A. Settari [4]. Capillary pressures, defined as the difference of phase pressures P_α , are described by the Parker approximate model with Van Genuchten parameters [5].

For numerical implementation of the above system an algorithm of the explicit type is proposed. On each j -th time level the next sequence of operations are fulfilled (starting from the initial and boundary conditions for primary variables P_w, S_w and S_n):

- Calculation of pressures P_n and P_g via P_w and capillary pressures;
- Calculation of phase densities from (9);
- Calculation of Darcy velocities from (8);
- Calculation of the term $(\rho_\alpha S_\alpha)$ on the next time level from (7) via the three-level explicit difference scheme;
- Calculation of P_w, S_w and S_n on the next time level solving the following system of three nonlinear algebraic equations in each point of the spatial grid (state equations (9) and the just obtained values $(\rho_\alpha S_\alpha)$ are used, P_{cnw} and P_{cgn} are capillary pressures):

$$\begin{cases} \rho_{0w} \left[1 + \beta_w (P_w^{j+1} - P_{0w}) \right] S_w^{j+1} = (\rho_w S_w)_i^{j+1} \\ \rho_{0n} \left[1 + \beta_n (P_w^{j+1} + P_{cnw} (S_w^{j+1}) - P_{0n}) \right] S_n^{j+1} = (\rho_n S_n)_i^{j+1} \\ \rho_{0g} \frac{(P_w^{j+1} + P_{cnw} (S_w^{j+1}) + P_{cgn} (1 - S_w^{j+1} - S_n^{j+1}))}{P_{0g}} \times \\ \times (1 - S_w^{j+1} - S_n^{j+1}) = (\rho_g S_g)_i^{j+1} \end{cases} \quad (11)$$

This system can be solved, for example, by Newton's method that takes only a few iterations.

- Data exchange at multiprocessor computing.

III. PARALLEL IMPLEMENTATION AND TEST PREDICTIONS

All computations have been performed on hybrid supercomputer K100 built in Keldysh Institute of Applied Mathematics (Moscow). The system consists of 64 nodes, each of them includes 2 six-core CPUs (Intel Xeon X5670, 2.93 GHz), 3 graphics accelerators (Nvidia Fermi C2050, 448 cores, 1.15 GHz, 2.5 GB of GDDR5) and 96 GB of DDR3 SDRAM. The original communication system named MVS-Express is constructed on the basis of PCI-Express and Infiniband with the transfer rate up to 700 MB/s, the latency is about 1.2 microseconds. The cluster peak performance is 100 TFLOPS.

A parallel software library was earlier developed by the

authors to simulate processes in the subsurface using supercomputers with graphics accelerators [3]. The kinetically-based model and corresponding explicit algorithm underlie the library. Due to the logical simplicity the algorithm is easily adapted to hybrid architectures of modern supercomputers.

The code is written in C/C++ in combination with CUDA and MPI. The library is oriented on solving problems in regular domains covered by Cartesian computational grids. It allows computations of 3D problems with double precision. Modular programming strategy is realized — the library consists of calculation, communication and control modules. Data partitioning (the geometrical parallelism principle) and data exchange on inner boundaries of sub-domains are chosen as the parallelization technique. Library routines can run on any number of CPU cores or GPUs from different nodes of a NUMA cluster. Approaches to exploit GPUs are discussed by the authors in papers [3], [6]. In particular it is pointed out that optimization of the access to various types of GPU memory is very important. In the current implementation cached memory has the preference, invariable parameters of the problem are loaded to the constant memory, the number of the global memory accesses is reduced, the register and local memory is used if possible, the shared and texture memory is not used.

The library is now supplemented with new computational modules for modeling three-phase fluid flows. Furthermore, the original procedure of automatic data partitioning among processing units (CPUs or GPUs) is incorporated into the library in order to provide the optimal load distribution on the basis of a priori estimation of the run time.

The computational domain can be divided into subdomains in one, two or three directions. The total run time of an application on a hybrid cluster consists of three main components: the time of actual numerical calculations, the time of data exchange between cluster's nodes and the time of copying data between the operative memory and the own GPU memory within one node. The share of each component depends as on the given computer system as on the configuration of domain partitioning. The algorithm of optimal domain partitioning is carefully described in [7].

During the current research some three-phase fluid flows in a homogeneous porous medium were predicted. Thus the process of three-phase fluid infiltration into the soil, namely redistribution of phases in a porous medium under the gravity influence, has been simulated. The computational domain is a parallelepiped with an impermeable boundary surface. This reservoir is filled with sand saturated by water, oil and air uniformly. The initial conditions are as follows:

$$\begin{aligned} S_w &= 0.4, \quad S_n = 0.35, \quad S_g = 0.25, \\ P_w &= P_{atmospheric} + \rho_w g h. \end{aligned} \quad (12)$$

The problem is quasi-one-dimensional therefore the obtained results can be presented as profiles of phase saturations at different time moments (see Fig.1 (a), (b)). The coordinate x is the distance from the reservoir top.

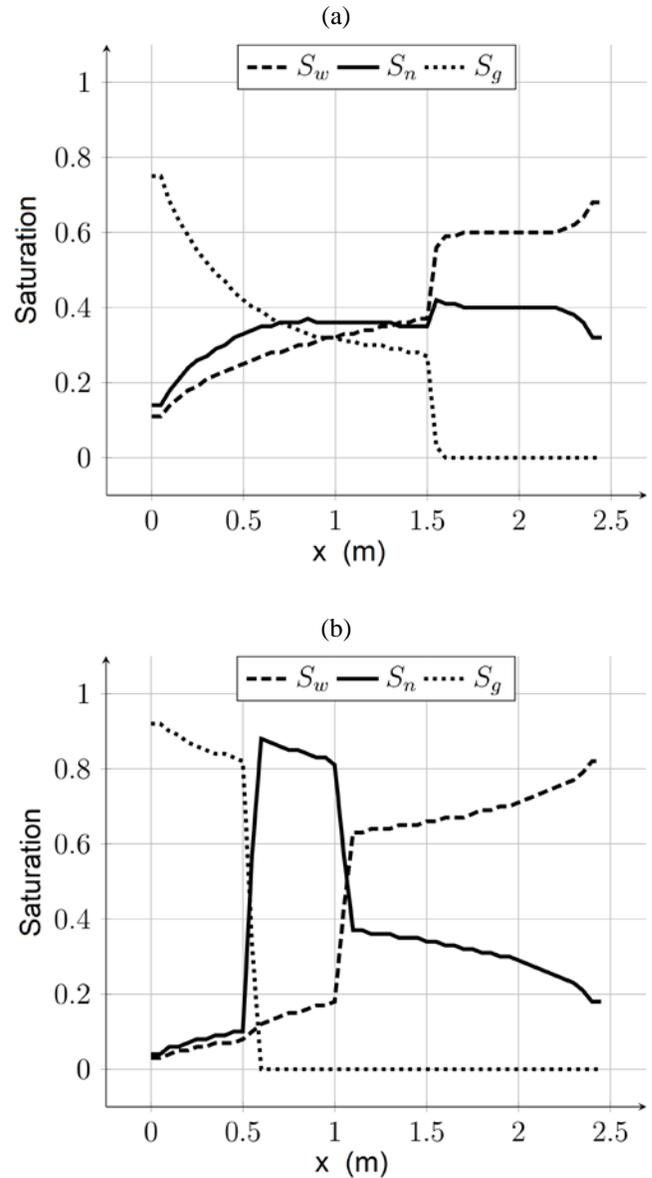


Fig.1 Profiles of saturations in the problem of phase redistribution at two different time moments: (a) - 83 minutes, (b) - 1500 minutes.

One can observe that under the gravity the water phase occupies mainly the lower part of the reservoir, the oil phase – basically the middle and slightly the lower part, and the gas phase – almost the whole upper part. Such distribution is physically correct. It is caused by the difference in phase densities and the given functions of relative phase permeability.

At the solution of the above test problem visible effect of the use of optimal domain partitioning is demonstrated. The computational grid size is $200 \times 200 \times 100 = 4$ million points. The number of operating CPU cores changes from one to a hundred and for each fixed number of processing units the own optimal domain distribution is chosen automatically. Strong scaling of the code is estimated when run times of 50 time steps are measured. One can see that the speedup (Fig.2 (a)) is

close to linear one and the efficiency (Fig.2 (b)) fluctuates near 90 per cent. Jumps of the efficiency occur due to different configurations of the domain distribution depending on the number of CPU cores.

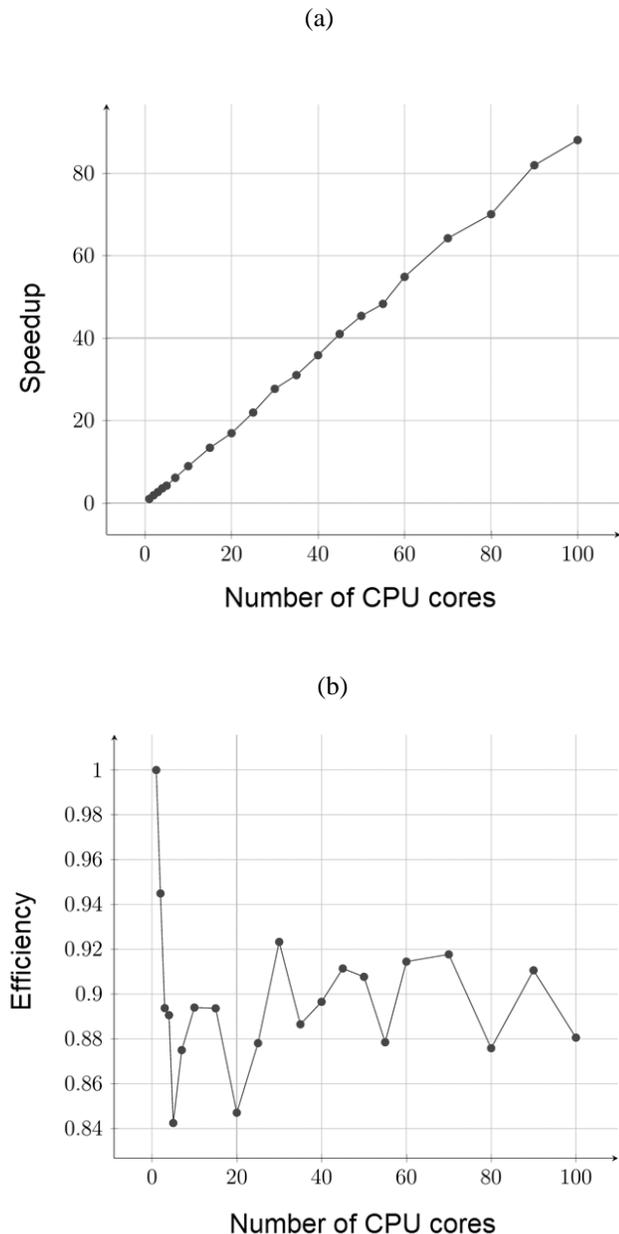


Fig.2 Strong scaling of the code for phase redistribution problem: (a) - speedup, (b) - efficiency.

A similar problem has been solved in the cubic domain. The process of water infiltration into an impermeable box through the hole situated in the centre of box's top is considered. Like the previous case the porous medium inside the box is initially saturated by water, oil and air uniformly. Equation (7) includes now a constant source of water. The speedup of computations on hybrid cluster K-100 was estimated for this problem solution. The code ran as on CPUs as on GPUs from different nodes of the cluster. The speedup of about 45 was achieved

when a GPU was compared with a CPU core, the computational grid size was 6.3 million points. The time needed for calculation of 100 time levels of the problem was measured: a CPU core took 30 hours, but a GPU – only 45 minutes for these computations. Weak scaling of the code was investigated at simultaneous increase of the number of computational grid points and employed GPUs. GPUs' run times were compared with run times of one CPU core at the same grid refinement. Thus the code runs about 270 times faster on six GPUs versus one CPU core at the computational grid size of 37.8 million points.

IV. CONCLUSION

The performed research confirms that in spite of difficulties of GPU-based supercomputer employment GPGPU computing is very promising for the solution of large-scale applied problems, in particular, for detailed description of flows in the subsurface.

In the nearest future the created approach will be complicated by allowing for heat transmission in order to simulate perspective thermal methods of oil recovery based on in-situ combustion. The presented investigation can be useful for the prediction, optimization and control of hydrocarbon recovery processes, for the realization of new technologies to increase the oil production rate.

REFERENCES

- [1] B. N. Chetverushkin, *Kinetic schemes and Quasi-Gas Dynamic system of equations*. Barcelona: CIMNE, 2008.
- [2] D. N. Morozov, M. A. Trapeznikova, B. N. Chetverushkin and N. G. Churbanova, "Application of Explicit Schemes for the Simulation of the Two Phase Filtration Process," *Mathematical Models and Computer Simulations*, vol. 4, no. 1, 2012, pp. 62–67.
- [3] D.N. Morozov, M.A. Trapeznikova, B.N. Chetverushkin, N.G. Churbanova, "Simulation of filtration problems on hybrid computer systems," *Mathematical Models and Computer Simulations*, vol. 5, no. 3, 2013, pp. 208-212.
- [4] K. Aziz, A. Settari, *Petroleum reservoir simulation*. London: Applied Science Publ. Lmt., 1979.
- [5] J.C. Parker, R.J. Lenhard, T. Kuppasami, "A parametric model for constitutive properties governing multiphase flow in porous media," *Water Resources Research*, vol. 23, no. 4, 1987, pp. 618-624.
- [6] M. Trapeznikova, B. Chetverushkin, N. Churbanova, D. Morozov, "Two-phase porous media flow simulation on a hybrid cluster," in *LSSC 2011, Lecture Notes in Computer Science*, vol. 7116, I. Lirkov, S. Margenov, and J. Wansiewski, Eds. Berlin: Springer, 2012, pp. 646-653.
- [7] M.A. Trapeznikova, N.G. Churbanova, A.A. Lyupa, D.N. Morozov, "Simulation of Multiphase Flows in the Subsurface on GPU-based Supercomputers," in *Parallel Computing: Accelerating Computational Science and Engineering (CSE), Advances in Parallel Computing*, vol. 25, M. Bader, A. Bode, H.-J. Bungartz, M. Gerndt, G.R. Joubert, F. Peters, Eds. Amsterdam: IOS Press, 2014, pp. 324-333.

Efficient distribution conversion algorithm in low power TRNGs for embedded security applications

Blerim Rexha, Dren Imeraj, Ehat Qerimi and Arbnor Halili

Abstract—The raw analog data generated from almost any noise source in nature has a normal distribution of values. Cryptography applications usually require true random number generators (TRNGs) to output random data streams that have a uniform distribution of values. Since TRNGs use a noise source to generate the random data, and that source usually has a normal distribution, the TRNG has to convert the distribution. If the TRNG is implemented in a low power device such as a microcontroller, the algorithm for distribution conversion needs to be lightweight and efficient in terms of using as much of the raw data as possible. Current market implementations convert the analog data into digital data by employing a comparator or a Schmitt trigger. This method wastes a large amount of random input data, lowering the throughput of the TRNG. This paper presents a novel algorithm that enables distribution conversion in low power devices. The low memory requirements and efficient processing make it suitable for implementation in microcontrollers or other low power cryptographic devices. The algorithm is also flexible, allowing for any size of noise samples.

Keywords—Algorithm, distribution, random, security, TRNG.

I. INTRODUCTION

RANDOM number generation for cryptography applications is usually a critical operation. The random data required is often used to form encryption keys, in this case, the random number generator must be practically unpredictable as described in [1]. Predicting the outcome of the random number generator (RNG) can lead to access of encrypted data by an unauthorized third party. An example of this is the exploitation of Dual Elliptic Curve Deterministic Random Bit Generator's flaw is presented in [2] by the NSA. In order to minimize the possibility of outcome prediction, critical cryptography applications employ true random number generators (TRNGs) when feasible. These RNGs use a physical noise signal as the source of the random data. Since the signal output from some noise sources is practically impossible to predict, these signals are ideal for use as RNGs.

This work was supported Ministry of Education and Science and Technology of Republic of Kosovo under contract signed in 11.12.2013 in Prishtina.

Blerim Rexha is professor and head of Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Prishtina, 10000 Prishtina, Kosovo (e-mail:blerim.rexha@uni-pr.edu).

Dren Imeraj is software developer and external associate to Faculty of Electrical and Computer Engineering, University of Prishtina, 10000 Prishtina, Kosovo (e-mail:dren.imeraj@uni-pr.edu).

Ehat Qerimi is software developer and external associate to Faculty of Electrical and Computer Engineering, University of Prishtina, 10000 Prishtina, Kosovo (e-mail:ehat.qerimi@uni-pr.edu).

Arbnor Halili is software developer and external associate to Faculty of Electrical and Computer Engineering, University of Prishtina, 10000 Prishtina, Kosovo (e-mail:arbnor.halili@uni-pr.edu).

Unfortunately, the distribution of the values in a noise source is usually not suitable for use in cryptography. Cryptography applications almost always require the data distribution to be uniform [3], meaning the probability that the RNG will generate any value in the value domain is always the same. On the other hand, many noise sources that are suitable for use in RNGs usually have a normal distribution according to [4]. The signal coming out of these sources oscillates around a specific point, i.e. 0V or some DC offset, and the possibility that some value X will be generated drops rapidly with the increase in magnitude of X. In order to use these data in cryptography applications, some means of converting the distribution from normal to uniform must be applied first.

II. TRADITIONAL RANDOM DATA HARVESTING METHODS

Converting data with a normal distribution to data with a uniform distribution can be easily achieved if data losses are tolerated. A simple way to achieve this conversion is by using a comparator with a threshold positioned at the DC offset of the noise signal as shown in [5]. Since the noise has a normal distribution, represented by a Gaussian curve as presented in Fig. 1, with the peak of the curve located at the DC offset, the probability that a certain value is located below the DC offset is the same as the probability that the value is located above the DC offset.

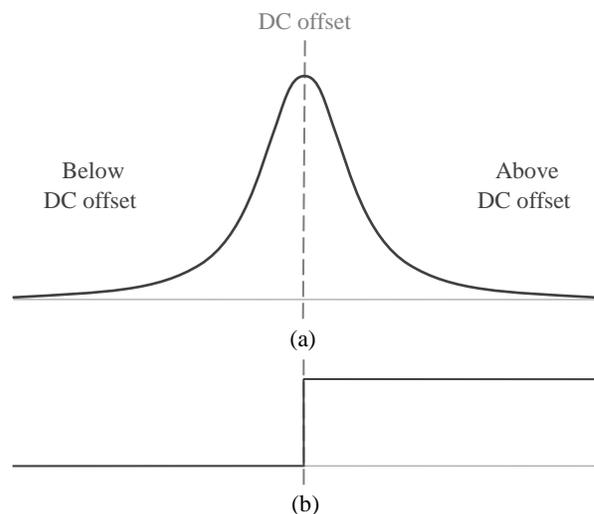


Fig. 1 Comparator output (b) depending on the noise signal input (a)

If the output of the comparator is represented as a binary value, i.e. binary 0 if the noise signal is below the DC offset and binary 1 if the noise signal is above the DC offset, these data can be packed into bytes. Since each bit in any generated byte has the same probability of being either binary 1 or binary 0 as any other bit, the distribution of the generated bytes will always be uniform.

The main problem with the solution above is the data losses during signal quantization as described in [6] and [7]. Since the noise signal is analog, each sample taken from the signal using the solution presented above contains more information than whether the signal is above or below the DC offset. This implies that the solution presented above is not efficient in terms of data throughput.

Another problem with the solution above is the difficulty modifying the noise signal before sampling the bits. Some noise sources do not originally have a normal distribution of values, therefore they require some operations to be applied in the noise signal in order to make them suitable for conversion. An example of this is the output of a reverse PN junction noise source.

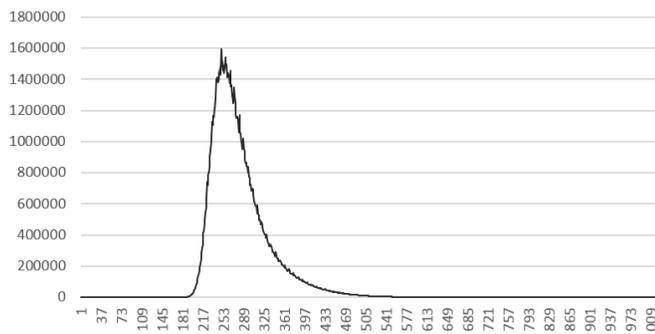


Fig. 2a The original distribution of the noise signal

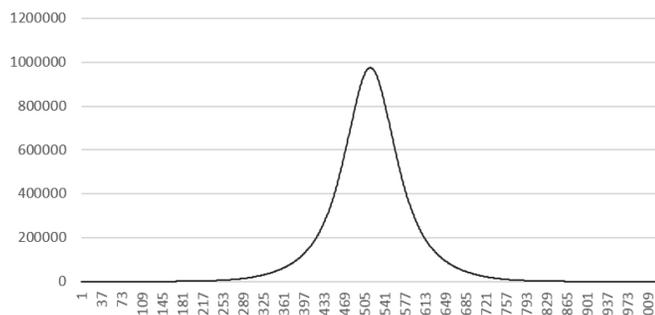


Fig. 2b The distribution of the values obtained by calculating the difference between two consecutive samples

In Fig. 2a is presented the original distribution of over 130 million samples taken directly from the noise source is shown. Clearly, the distribution in Fig. 2a is far from being a normal distribution. Fig. 2b shows the distribution of the values obtained by calculating the difference of two consecutive samples, X_i and X_{i-1} , from more than 130 million samples used previously. This figure shows a clear normal distribution. If a signal with the distribution, as presented in Fig. 2a is fed into a comparator, no matter where the threshold is specified, the probability of getting a binary 1 and the probability of getting a binary 0 out of the comparator will not be the same since the distribution is not symmetrical around the DC offset.

To overcome the problems described above, this paper presents an algorithm that uses samples of the noise signal quantized using an analog to digital converter. This way, the device running the algorithm can use more information from each sample as well as be able to pre-process the samples in

order to alter the distribution of the input signal before feeding the samples to the distribution conversion algorithm

III. GENERATING RAW DATA

In order to test the algorithm presented in this paper, special hardware has been designed and implemented. The noise source used in this implementation is a reverse biased PN junction as described in [8]. This noise source is truly random since its signal is a result of two non-deterministic physical phenomena, quantum tunneling and avalanche multiplication [9], [10]. This noise source has been chosen for its good noise quality, high frequency and high stability.

The noise generation circuit consists of a BJT transistor with its base-emitter terminals reverse biased. Under high enough reverse voltage on those terminals, the current flowing through this transistor will become noisy due to the effects mentioned above. This current is amplified by another BJT transistor which is later AC coupled. The AC coupling will remove the DC offset generated by the amplifying BJT transistor since the magnitude of that DC offset is too high compared to the magnitude of the noise signal. To make the noise signal suitable for the analog to digital converter, a new DC offset is added on the noise signal. Since the impedance of the DC offset generation resistors and the AC coupling capacitor is high, the analog to digital converter may alter the noise signal while sampling it as shown in [11]. To overcome this issue, a buffer is added between the analog to digital converter and the DC offset generator resistors. This buffer consists of an operational amplifier configured a unity gain buffer. The reason for using a unity gain configuration on the operational amplifier is because of the gain-bandwidth product associated with the operational amplifier. As seen in [12], increasing the gain of the operational amplifier reduces its bandwidth. The problem introduced by the lack of amplification is that the noise signal level is low. To overcome this problem, a suitable reference voltage for the analog to digital converter is used.

Fig. 3 presents a schematic diagram of the implemented noise generation circuit. The reverse bias voltage is generated by a boost converter that converts 5V to 30V. Feeding 30V to the noise generation and amplification stage will generate about 400mV peak AC coupled. The DC offset generation stage will generate an offset of 550mV, which will make the signal suitable for the analog to digital converter that uses a 1.1V reference voltage. An important thing to consider is the values of the AC coupling capacitor and the DC offset generation resistors. Since this stage acts as a high pass filter, it is important to choose the values properly so that important low frequency components of the noise signal are not attenuated.

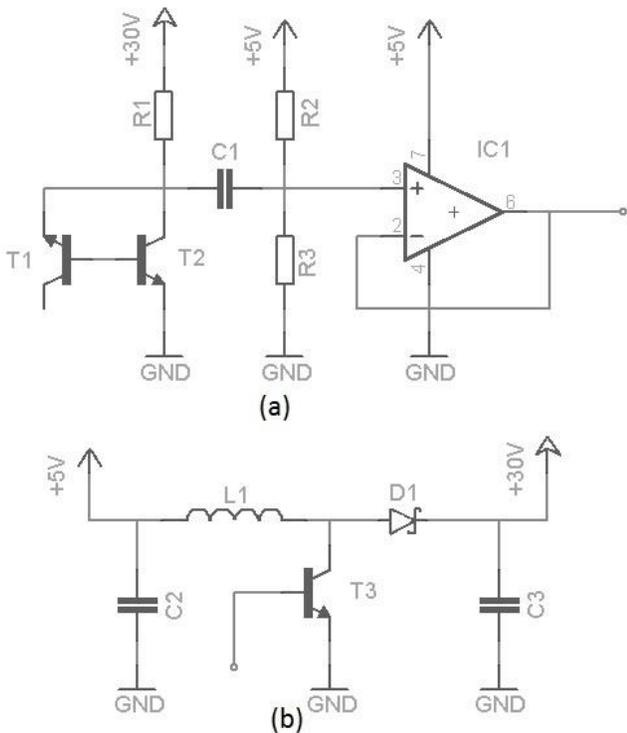


Fig. 3 – Noise generator circuit (a) and boost converter (b)

IV. SHIFTING TO UNIFORM DISTRIBUTION

Suppose a set consisting of a finite number of 10 bit randomly generated values with a normal distribution. Also suppose that this set is represented as a binary matrix, with its rows containing the sample's bits and the columns showing the samples themselves, as presented in Table 1

Table I- An example of a binary matrix of samples

MSB (9)	0	1	0	0	1	0	1	1	0	0
8	1	1	0	1	0	1	1	1	0	1
7	1	0	1	1	1	1	0	1	0	1
6	1	1	1	1	0	1	1	1	1	0
5	1	1	0	1	1	0	0	0	0	1
4	1	0	1	0	1	1	0	0	1	1
3	1	0	1	1	0	1	1	0	0	0
2	0	1	0	1	0	0	1	1	0	0
1	1	0	1	0	0	0	1	0	1	1
LSB (0)	1	1	1	0	1	0	1	0	1	0

If one calculates the average time it takes a bit position (on a specific row) to change its state, it can be concluded that the time interval increases exponentially going from the least significant bit (LSB) to the most significant bit (MSB). In order for this data to be used as random data, the average time it takes the LSB to change its state must not be higher than 1 sample, otherwise each two following samples would have the same value.

Since the LSB takes 1 sample to change its state, the time it takes other bit positions to change their state can also be represented in samples. In order to shift to uniform

distribution, each bit position must have the same average time it takes to change its state, 1 sample.

The goal is to reconstruct a new set of N-bit values with a uniform distribution. Since it is known that the LSB takes an average of 1 sample to change its state, each LSB can be used to construct the new set of N-bit values. But this is not true for the other bit positions. Since other bit positions take an average of more than 1 sample to change their state, these bit positions must wait before they are used. The waiting time depends on their average time it takes them to change the state. If a bit position takes an average of 2 samples to change its state, only one in two samples from the original set can pass this bit to the new set. In the meantime, the bit positions that are waiting can be XORed.

By using the bit positions not more than their average time needed to change their state in the original set, it is ensured that each bit in the new set has an average time of state change of 1 sample

V. ALGORITHM IMPLEMENTATION

In the Fig. 4 is presented the C pseudo code, in accordance to the algorithm definition in Section IV. The array holding the *Thresholds* can be experimentally determined using the algorithm defined in Section VI.

```

ConvertToUniform (Data, DataLength):
    n = 0
    XORSample = 0
    XOR_Counts = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0}
    Thresholds = {1, 2, 4, 8, 16, 32, 64, 128, 256, 512}

    WHILE n < DataLength
        XORSample = XORSample XOR Data[n]
        n = n + 1

        FOR i = 0 TO 9
            XOR_Counts[i] = XOR_Counts[i] + 1

            IF XOR_Counts[i] IS EQUAL TO Thresholds[i]
                IF BIT i OF XORSample IS 0
                    ADD BIT 0 TO OUTPUT
                ELSE
                    ADD BIT 1 TO OUTPUT
                ENDF
            XOR_Counts[i] = 0

            IF i IS EQUAL TO 9
                XORSample = 0
            ENDF
        ENDFOR
    ENDWHILE
    RETURN
    
```

Fig. 4 – C pseudo code for conversion algorithm

The *InsertBitToOutput(BitValue)* function inserts the bit given as a parameter to the output data. At this point, the data constructed by this function has a uniform distribution since the bits given as a parameter the function have passed through the distribution conversion function.

VI. DETERMINING THE THRESHOLDS

In order to make use of most of the data generated by the noise source, the Thresholds array of the algorithm must be determined based on the characteristics of the noise source. Index N of the Thresholds array holds the average number of samples that need to be read before the state of bit position N changes. Since the characteristics of each noise source differ, in order to achieve the best results the Thresholds array must be determined experimentally. This can be done once, if the characteristics of the noise source do not drift, or the values of the Thresholds array can be updated each time a new sample is obtained from the noise source. The first method requires less computational resources but can lead to errors in the generated random data, while the second method is more resource intensive but offers better reliability.

Since the Thresholds array holds the number of samples that need to be read before a specific bit position can be used, the array can be constructed by calculating the number of bit flips on a large set of samples and dividing the number of samples in the set by the number of bit flips.

The C pseudo code, as presented in Fig. 5, can be used to determine the *Thresholds* array for a given sample set.

```

CalculateThresholdsArray(SampleSet, SampleCount,
                        Thresholds, SampleSize):
FOR n = 0 TO SampleSize - 1
    IF BIT n OF SampleSet[0] IS 0
        LastBit = 0
    ELSE
        LastBit = 1
    ENDIF
    FlipCount = 0
FOR i = 1 TO SampleCount - 1
    IF BIT n OF SampleSet[i] IS 0
        CurrentBit = 0
    ELSE
        CurrentBit = 1
    ENDIF
    IF (LastBit IS 1 AND CurrentBit IS 0) OR
        (LastBit IS 0 AND CurrentBit IS 1)
        FlipCount = FlipCount + 1
        LastBit = CurrentBit
    ENDIF
ENDFOR
Thresholds[n] = SampleCount / FlipCount
ENDFOR
RETURN
  
```

Fig. 5 – C pseudo code for calculating the threshold array

The sample set fed to the function as defined in Fig. 5 must be as large as possible in order to get good results.

VII. QUALITY OF RESULTS

The quality of the algorithm is measured based on the performance and the algorithm in terms of data throughput and based on the quality of the generated random data.

The throughput of the algorithm is based on the Thresholds array. The number of uniformly distributed bits that can be obtained from an N bit normally distributed input value, as presented in (1).

$$M = \sum_{i=0}^N \frac{1}{Thresholds[i]} \quad (1)$$

Using (1), it can be concluded that the *Thresholds* array in Section V has an output of almost 2 uniformly distributed bits for each normally distributed input sample. Since the threshold for the LSB will always be 1 (otherwise the data cannot be used as random data, as described in Section IV), and the thresholds for the other bits are always finite, the effective output of the algorithm is always more than 1 bits for each sample. Compared to the traditional method of extracting uniformly distributed data using a comparator or a Schmitt trigger **Error! Reference source not found.**, the algorithm offers higher throughput since the traditional methods offer a throughput of 1 bit for each sample.

The quality of the generated data is measured using a set of randomness testing tools offered by the CrypTool software, most of which are proposed by NIST in [13] and Intel in [14]. In the Table II are presented the tests performed of 128KB of data generated by the algorithm:

Table II Randomness test results

Test	Max. test value	Test value
Frequency test	3.841	0.623
Poker test	14.07	5.423
Runs test	9.488	4.217
Serial test	5.991	2.167

Another test, also proposed in [13], which indicates missing patterns in the generated data is the entropy. The entropy indicates the level of “chaos” in the data. If the generated data is organized in bytes (8 bit words), the resulting entropy of the 128KB of data used for the tests, as presented in Table 2, using CrypTool the value of entropy comes out 7.99. With such a high entropy, compression algorithms would not be able to find any patterns in the data, making it impossible to compress the generated random data.

VIII. CONCLUSIONS

The algorithm presented on this paper increases the distribution conversion performance compared to the traditional method of using a comparator or a Schmitt trigger.

The implementation presented in section V is lightweight and can easily be implemented in microcontrollers or other embedded applications. The same microcontroller can be also

used to sample the analog noise signal which makes the TRNG cost effective and small in size.

The algorithm can be dynamically adopted to obtain as much data from the noise source as possible. This is done by updating the *Thresholds* array every time a new sample is obtained.

The implementation of the distribution conversion in software also offers the flexibility of specifying the output data format, as well as the distribution itself. The user can specify whether the distribution should be converted or not.

In order to test the algorithm, a special electrical circuit has been built. This circuit uses the reverse biased Emitter-Base junction of a BJT transistor to generate a noise signal. This signal is then fed to the analog to digital converter of a microcontroller which samples the noise signal and feeds it to the algorithm. Prior to developing the algorithm, the circuit was used to create a cache of 130 million 10 bit samples of the noise. These data were used to test the algorithm during the development. As described in Section VII, the algorithm passes the frequency test, poker test, runs test and serial test. It is important to note that the data in generated by the algorithm has an entropy of 7.99 when packed in eight bit words.

In conclusion, the algorithm presented in this paper is suitable for embedded security applications that require a TRNG with a relatively high data throughput. The noise signal can also be different from pure Gaussian and then pre-processed in the microcontroller before being fed to the algorithm.

The future work remains to add this efficient TRNG as web services for other application, interested for reliable random data using any well proven encryption protocols for securing data transfer.

REFERENCES

- [1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
- [2] W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
- [3] H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.
- [4] B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.
- [5] E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," *IEEE Trans. Antennas Propagat.*, to be published.
- [6] J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.
- [7] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
- [8] Y. Yoroazu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," *IEEE Transl. J. Magn.Jpn.*, vol. 2, Aug. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics Japan*, 1982, p. 301].
- [9] M. Young, *The Technical Writers Handbook*. Mill Valley, CA: University Science, 1989.
- [10] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility (Periodical style)," *IEEE Trans. Electron Devices*, vol. ED-11, pp. 34–39, Jan. 1959.

- [11] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Trans. Neural Networks*, vol. 4, pp. 570–578, July 1993.
- [12] R. W. Lucky, "Automatic equalization for digital communication," *Bell Syst. Tech. J.*, vol. 44, no. 4, pp. 547–588, Apr. 1965.
- [13] S. P. Bingulac, "On the compatibility of adaptive controllers (Published Conference Proceedings style)," in *Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory*, New York, 1994, pp. 8–16.
- [14] G. R. Faulhaber, "Design of service systems with priority reservation," in *Conf. Rec. 1995 IEEE Int. Conf. Communications*, pp. 3–8.
- [15]



Blerim Rexha was born on May 1st, 1970 in Prishtina, Kosovo. He graduated with distinguished mark from University of Prishtina, Faculty of Electrical and Computer Engineering. He received the Ph.D. from Vienna University of Technology, Institute of Computer Technology in 2003 in field of data security, smart cards and online security.

From year 2000 till 2007 he worked for Siemens AG, division of Program and System Engineering in Vienna starting as a software developer and finishing as project manager. During his career in Siemens he managed many projects such as Austrian patient "e-Card", software for German Telecom, application for digital signature, etc.

Prof. Rexha has published many scientific papers in its field of interest and also he is owner of two patents for securing online transaction. Currently he is head of computer engineering department at the Prishtina University, Faculty of Electrical and Computer Engineering where he teaches the courses about data security, network security and software engineering.

Successive elimination algorithm for truncated gray-coded bitplane matching based motion estimation

Ilseung Kim, and Jechang Jeong

Abstract—In this paper, a successive elimination algorithm for truncated gray-coded bitplane matching (TGCBPM) based motion estimation (ME) is proposed. After calculating the lower bound for TGCBPM, we can substantially reduce the computational complexity by skipping the impossible candidate. Experimental results show that this proposed algorithm has an outstanding performance in terms of the computational complexity although the motion accuracy of the proposed algorithm is the same as that of the full search TGCBPM.

Keywords— Video coding, motion estimation, full-search, truncated gray-coded bit-plane matching (TGCBPM)

I. INTRODUCTION

Motion estimation (ME) plays a critical role in the video compression process, since it considerably contributes coding efficiency by reducing possible temporal redundancies between frames. The full search-based sum of absolute differences (FS-SAD) or sum of squared differences (FS-SSD) methods can be the optimal estimation of the motion, which conducts exhaustive searches within its search range in order to find the minimum matching error based on sum of absolute differences. However, the FS-SAD method is not suitable for real-time applications because it requires excessive computation.

In order to alleviate this problem, many approaches have been proposed. Most of the techniques try to minimize the whole calculations of the matching criterion [1]-[4] or searching points and successive elimination algorithm (SEA) is one of those [5]. In SEA, if the pre-calculated lower bound of the block matching error is larger than the up-to-date minimum distortion, the block matching process of the block will be skipped safely. With this process, a number of the matching computations are

This research was supported by the MSIP(Ministry of Science, ICT & Future Planning), Korea, under the project for technical development of information communication & broadcasting supervised by IITP(Institute for Information & Communications Technology Promotion)(IITP 2015-B0101-15-1377).

Ilseung Kim is with Hanyang University, Seoul, Korea. He is now with the Department of Electronics and Computer Engineering, Seoul, Korea. (e-mail: ghanjang@gmail.com).

Jechang Jeong is with Hanyang University, Seoul, Korea. He is now with the Department of Electronics and Computer Engineering, Seoul, Korea. (phone: 82-2-2220-4369; e-mail: jjeong@hanyang.ac.kr).

reduced, while the ME accuracy of the SEA is the same as that of the full search algorithm.

Some algorithms have been proposed with the different approaches that utilize different matching criteria from the conventional SAD or SSD, which exploit Boolean operations in order to speed up the calculation of the matching criterion itself [6]-[10]. These techniques are called binary block motion estimation or bit-plane matching (BPM)-based MEs. These techniques have two main advantages compared to the typical ME algorithms based on SAD or SSD: fast computation of the matching criterion using Boolean operation and reduced memory bandwidth.

In [6], bit planes for motion estimation are generated by comparing the original image frame against the multi-bandpass filtered frame, whose method is called one-bit transform (1BT)-based ME. In [7], a multiplication-free 1BT (MF1BT) is proposed, which exploits a multiplication-free bandpass filter kernel in order to generate bit planes. Two-bit transform (2BT) and constrained 1BT (C-1BT) were suggested to improve the motion accuracy of 1BT-based ME approaches [8], [9]. In [8], two bit-planes are constructed employing basic statistical information such as mean and standard deviation. In [9], a bit-plane, called a constrained bit-plane, is added to refine the performance of 1BT-based ME. In [10], the truncated gray coded bit-plane matching (TGVBPM) based ME was proposed. This method switches the original frames' pixels into gray-coded ones, chooses some bit planes from the most significant bit (MSB), and conducts the block matching process using the chosen bit planes. The authors of [10] recommend that 3 bit-planes are good enough to improve performance of previously presented BPM-based ME methods.

BPM-based ME methods take advantage of fast computation of their criteria, but they still exploit FS strategy for all possible candidate in the search range. By using the fact, several algorithms were proposed to further reduce the computational complexity of the BPM-based ME algorithm by exploiting early termination scheme [11]-[12] and combining search range adjustment method [13]. Note that the ME accuracy of those algorithms are not the same as that of the FS-BPM-based ME methods.

SEA method is one of the techniques which effectively reduces the complexity of BPM-based ME without any loss [14] – [16]. The SEA for 1BT was derived based on the triangle

inequality to remove the impossible candidates in [14]. In [15], the authors derived the lower bound of the block matching error of 2BT by using set theory and triangle inequality. Recently, SEA for C1BT was proposed [16]. Because the matching criterion of C1BT does not satisfy the metric conditions, the authors derive the lower bound of the block matching error using the Bonferroni inequality.

In this paper, SEA for TGCBPM is proposed. The proposed algorithm analyzes the TGCBPM matching criterion and derives the lower bounds of it based on mathematical techniques in order to discard the impossible candidates and save computations significantly. Experimental results demonstrate that the proposed algorithm reduces computational complexity substantially while maintaining the entirely same ME accuracy.

The rest of this paper is organized as follows. In section II, we review the previous SEA techniques for 1BT and 2BT, and C1BT. Section III presents our proposed algorithm. Experimental results are provided in Section IV. The conclusion is set forth in Section V.

II. PREVIOUS WORKS

In this Section, the previous works of SEA algorithms for BPM based ME methods are reviewed including SEAs for 1BT and 2BT, because some results in previous works are some of the bases of the proposed algorithm.

A. Successive Elimination Algorithm for 1BT

The derivation of the SEA for 1BT in [14] starts from the basic triangle inequality for binary values. Let b_1 and b_2 be any two one-bit binary value, and then the following inequality holds:

$$|b_1 - b_2| \leq b_1 \oplus b_2 \quad (1)$$

where \oplus denotes the Boolean X-OR operation.

Stretching the meaning of this result in terms of the matching measure of 1BT, i.e. $NNMP_{1BT}$, we can derive the following result:

$$\begin{aligned} NNMP_{1BT}(m, n) &= \sum_{i,j=0}^{N-1} |B^t(i, j) \oplus B^{t-1}(i+m, j+n)| \\ &\geq \sum_{i,j=0}^{N-1} |B^t(i, j) - B^{t-1}(i+m, j+n)| \\ &\geq \left| \sum_{i,j=0}^{N-1} B^t(i, j) - B^{t-1}(i+m, j+n) \right| \end{aligned} \quad (2)$$

where B^t denotes the one-bit representation of the current frame, B^{t-1} denotes that of the reference frame, (i, j) represent the search points in the search range $[0, N-1]$, and (m, n) represent displacement of a motion vector, respectively. If the sum norm of the block is not satisfied with (2), the search is not

performed at the point. The calculation of the sum norm can be done efficiently by using the method in [5].

B. Successive Elimination Algorithm for 2BT

Let \mathbf{B} be a vector as binary sequences $b(i)$ ($0 \leq i \leq N-1$). The authors in [15] introduce the concept of the Hamming weight in order to concisely express the sum norm of the blocks and utilize the set theory. The relationship between sum norm of $b(i)$ and the Hamming weight of \mathbf{B} can be driven with the following equality:

$$\sum_{i=0}^{N-1} b(i) = w_H(\mathbf{B}) \quad (3)$$

where $w_H(\cdot)$ denotes the Hamming weight which is the number of nonzero components.

By using the concept of the set theory, we obtain the following inequality:

$$w_H(\mathbf{x} \parallel \mathbf{y}) \geq \max(w_H(\mathbf{x}), w_H(\mathbf{y})) \quad (4)$$

where \parallel denotes the Boolean OR operation and \mathbf{x} and \mathbf{y} represent arbitrary two vectors.

Matching criterion of 2BT is given as

$$\begin{aligned} NNMP_{2BT}(m, n) &= \sum_{i,j=0}^{N-1} \{B_1^t(i, j) \oplus B_1^{t-1}(i+m, j+n)\} \\ &\parallel \{B_2^t(i, j) \oplus B_2^{t-1}(i+m, j+n)\} \end{aligned} \quad (5)$$

where $B_{1,2}^t$ and $B_{1,2}^{t-1}$ are the two-bit representations of current and reference frames, (i, j) represents the search points in the search range $[0, N-1]$, (m, n) represents displacement of a motion vector, the motion block size is $N \times N$ and $-s \leq m, n \leq s$ is the search range.

The result of the SEA for 2BT can be obtained by using the results of (2) and (4) as follows:

$$\begin{aligned} NNMP_{2BT}(m, n) &\geq \max\left(\left|w_H(\mathbf{B}_1^t)\right| - \left|w_H(\mathbf{B}_{1,mm}^{t-1})\right|, \left|w_H(\mathbf{B}_2^t)\right| - \left|w_H(\mathbf{B}_{2,mm}^{t-1})\right|\right) \end{aligned} \quad (6)$$

where we identify a vector $\mathbf{B}_{1,2}^t$ as $B_{1,2}^t(i, j)$ and a vector $\mathbf{B}_{1,2,mm}^{t-1}$ as $B_{1,2}^{t-1}(i+m, j+n)$.

III. PROPOSED ALGORITHM

In this section, the derivation of the lower bound for TGCBPM matching criterion using the result in section II is presented. In order to derive the SEA for TGCBPM-based ME, the analysis of the TGCBPM-based ME algorithm is necessary.

A pixel value that is quantized to 2^K grey levels can be represented as:

$$f^t(x, y) = a_{K-1}2^{K-1} + a_{K-2}2^{K-2} + \dots + a_12^1 + a_02^0 \quad (7)$$

where a_k coefficients represent the natural binary code and take only binary values and K denotes the bit-depth for the pixel. The gray-coded version of a pixel value can be obtained from its natural binary codes with the following equations:

$$\begin{cases} g_{K-1} = a_{K-1} \\ g_k = a_k \oplus a_{k+1} \end{cases}, 0 \leq k \leq K-2 \quad (8)$$

The matching criterion for the TGCBPM-based ME, called the correlation metric CM_{TGC} , is given as

$$\begin{aligned} & CM_{TGC}(m, n) \\ &= \sum_{i,j=0}^{N-1} \sum_{k=NTB}^{K-1} 2^{K-NTB} \{g_k^t(i, j) \oplus g_k^{t-1}(i+m, j+n)\} \quad (9) \\ & -s \leq m, n \leq s-1 \end{aligned}$$

where (m, n) , s , k , NTB , and g_k^t denote the candidate displacement and search range, bit level, the number of truncated bit, and k th level of the binary pixel value with gray-coded version, respectively. TGCBPM based ME method makes only use of the highest $(K-NTB)$ bit-planes to compute the matching process. Note that, higher order bit-planes have higher weight by a scaling factor of 2^{K-NTB} . The ME performance of [10] depends on NTB and the authors in [10] suggested the value of NTB as 5, which means only 3 bits from the MSB are used for the block matching process.

The purpose of SEA is that the pre-calculated lower bound of the block matching error is used to skip the impossible candidate that makes the ME process faster. The performance of SEA depends on how fast the lower bound is calculated. The algorithms in [14]–[16] make the fast computation of the lower bound of the matching error by using mathematical techniques suggested in (1)~(6), which contributes to increasing the overall ME performance in terms of computational time.

However, it is not appropriate to directly apply the result from the previous works in case of TGCBPM. Applying the inequality (1) into the (7), we get the following inequality:

$$\begin{aligned} & CM_{TGC}(m, n) \\ &= \sum_{i,j=0}^{N-1} \sum_{k=NTB}^{K-1} 2^{K-NTB} \{g_k^t(i, j) \oplus g_k^{t-1}(i+m, j+n)\} \\ &\geq \sum_{k=NTB}^{K-1} 2^{K-NTB} \cdot \sum_{i,j=0}^{N-1} |g_k^t(i, j) - g_k^{t-1}(i+m, j+n)| \\ &\geq \sum_{k=NTB}^{K-1} 2^{K-NTB} \cdot \left| \sum_{i,j=0}^{N-1} g_k^t(i, j) - \sum_{i,j=0}^{N-1} g_k^{t-1}(i+m, j+n) \right|. \end{aligned} \quad (10)$$

The last part of (8) can be the lower bound of the TGCBPM matching error in order to apply the SEA method with the results from the previous works. Of course, a lot of candidates point are going to be skipped without performing the block matching process with the result of (8). However, the process to obtain the lower bound is not efficient, because several sum-norm must be calculated depending on the user parameter NTB . Assume that NTB is set as 5, as recommended, then the process calculating the sum norms requires six times only for one candidate search point. Note that the number of sum norms to be calculated depends on the value of NTB .

Note that the gray-code is a well-known code whose mapping function can be easily obtained without bit-transforming, once the length of the code is determined. In order to utilize the fact mentioned above, we define a parameter G^t as follow:

$$G^t(i, j) \equiv \sum_{k=NTB}^{K-1} 2^{K-NTB} g_k^t(i, j) \quad (11)$$

, which is the integer value of gray-code whose length is determined by the given value of the parameters NTB , K and the input pixel value, so the calculation of G^t can be implemented by look-up table.

Applying (1) and (11) into (9), we obtain the main result of the proposed algorithm as follow:

$$CM_{TGC}(m, n) \geq \left| \sum_{i,j=0}^{N-1} G^t(i, j) - \sum_{i,j=0}^{N-1} G^t(i+m, j+n) \right| \quad (12)$$

Once the value of G^t is obtained in (12), we have only two sum norms left and the number of sum norms does not depend on the value of NTB . In the search process, the right part of (12) is compared with the CM_{\min} which is the up-to-date minimum CM_{TGC} in the search process. We calculate the CM_{TGC} of (9) only if the condition of (12) is satisfied. Otherwise, we skip this search position, move on to the next search position, and save the unnecessary operations. Note that the calculation of sum norms can be done efficiently using the method in [5].

Table I Average PSNR performance (dB) comparison for the algorithms

Sequences	1BT	C1BT	2BT	TGCBPM	Proposed
Akiyo	41.69	42.05	41.88	41.90	41.90
Bus	24.51	25.04	24.91	25.34	25.34
Container	37.29	37.98	37.78	37.49	37.49
Dancer	27.92	28.55	28.84	29.95	29.95
Football	22.89	23.20	23.16	23.77	23.77
Foreman	29.80	30.35	30.09	30.75	30.75
Hall	32.85	34.30	33.64	34.12	34.12
News	31.71	33.65	33.54	34.42	34.42
Stefan	22.92	23.09	23.22	23.50	23.50
Table	29.85	30.53	30.24	30.74	30.74
Average	30.14	30.88	30.73	31.20	31.20

Table II Average reduction ratio of the calculated points (%)

Akiyo	96.13
Bus	54.08
Container	81.01
Dancer	86.02
Football	50.49
Foreman	73.69
Hall	86.94
News	91.37
Stefan	54.01
Table	68.66
Average	74.24

IV. EXPERIMENTAL RESULTS

The proposed algorithm is simulated with various video sequences: Akiyo, Bus, Container, Dancer, Football, Foreman, Hall, News, Stefan, and Table. All implemented block matching algorithms are programmed by Visual C++. The motion block size is 16 x 16 pixels and the search range is 33 x 33. The format of video sequence that we used is CIF (352 x 288) and only forward prediction is used. All the searching processes are performed in spiral order. The parameter NTB is set as 5, as recommended in the reference paper.

Table I shows the average PSNR (dB) of the BPM-based ME algorithms and the proposed algorithm. As shown in Table I, the PSNR of TGCBPM based ME is better than that of the other BPM-based ME algorithms. In terms of the PSNR point of view, we can conclude the PSNR performance of the proposed

algorithm is totally same as that of the TGCBPM-based ME algorithm from the data of the results in Table I and the mathematical derivation in section III.

In order to compare the performance of the proposed algorithm with that of TGCBPM based ME, we calculate the average reduction ratio of the calculated points. The performance gains of the proposed algorithm somewhat differ depending on the sequences. The reduction ratio of the sequences that have static motion, such as Akiyo and Container, is greater than that of the sequences that have complex motion, such as Football and Stefan. On the average, the performance gain of the proposed algorithm is 74.24% over FS-TGCBPM based ME in terms of the complexity point of view. Note that the average PSNR performance of the proposed algorithm and the TGCBPM based ME is the same.

V. CONCLUSIONS

In this paper, a SEA for TGCBPM based ME has been proposed to improve the performance of the TGCBPM based ME in terms of the computational complexity. The proposed algorithm suggests the efficient way to calculate the lower bound for TGCBPM matching criterion and to eliminate the impossible candidates earlier so that the computational complexity of the TGCBPM is reduced significantly. Experimental results show that the computational complexity has been reduced substantially, while the PSNR of the proposed algorithm is the same as that of the TGCBPM based ME.

REFERENCES

- [1] R. Li, B. Zeng, and M. Liou, "A new three-step search algorithm for block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 4, pp 438-442, Aug. 1994.
- [2] S. Zhu and K. Ma, "A new diamond search algorithm for fast block matching motion estimation," in *Proc. ICICS*, Singapore, vol. 1, pp 292-296, 1997.
- [3] C. Cheung and L. Po, "Normalized partial distortion search algorithm for block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 3, pp 417-422, Apr. 2000.
- [4] C. Cheung and L. Po, "Adjustable partial distortion search algorithm for block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp 100-110, Jan. 2003.
- [5] W. Li and E. Salari, "Successive elimination algorithm for motion estimation," *IEEE Trans. Image Process.*, vol. 4, no. 1, Jan. 1995.
- [6] B. Natarajan, V. Bhaskaran, and K. Konstantinides, "Low complexity block-based motion estimation via one-bit transforms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 4, Aug. 1997.
- [7] Sarp Ertürk, "Multiplication-free one-bit transform for low-complexity block-based motion estimation," *IEEE Trans. Signal Processing Letter.*, vol. 14, no. 2, Feb. 2007.
- [8] Alp Ertürk and Sarp Ertürk, "Two-bit transform for binary block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 7, pp 938-946, Jul. 2005.
- [9] Oguzhan Urhan and Sarp Ertürk, "Constrained one-bit transform for low complexity block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 4, pp 478-482, Apr. 2007.
- [10] Anil Çelebi, Orhan Akbulut, Oguzhan Urhan, and Sarp Ertürk, "Truncated gray-coded bit-plane matching based motion estimation and its hardware architecture," *IEEE Trans. Consum. Electron.*, vol. 55, no. 3, pp 1530-1536, Aug. 2009.

- [11] H. Lee and J. Jeong, "Early termination scheme for binary block motion estimation," *IEEE Trans. Consum. Electron.*, vol. 53, no. 4, pp 1682-1686, Nov. 2007.
- [12] H. Lee, S. Jin, and J. Jeong, "Early termination scheme for 2BT block motion estimation," *Electron. Lett.*, vol. 45, no. 8, pp 403-405, Apr. 2009.
- [13] I. Kim, J. Kim, G. Jeon, and J. Jeong, "Low-complexity block-based motion estimation algorithm using adaptive search range adjustment," *Opt. Eng.*, vol. 51, no. 6, 067010, Jun. 2012.
- [14] Y. Wang and G. Tu, "Successive elimination algorithm for binary block matching motion estimation," *Electron Lett.*, vol. 36, no. 24, pp 2007-2008, Nov. 2000.
- [15] C. Choi and J. Jeong, "Successive elimination algorithm for two-bit transform motion estimation," *IEICE ELEX*, vol. 7, no. 10, pp. 684-690, May 2010.
- [16] C. Choi and J. Jeong, "Successive elimination algorithm for constrained one-bit transform based motion estimation using the Bonferroni inequality," *IEEE Proc. Lett.*, vol. 21, no. 10, pp. 1260-1264, Oct. 2014.

Improving programming courses using aptitude testing and learning styles

Eva Milková, Karel Petránek

Abstract The area of software development has passed a rapid expansion. An essential part of studies at faculties preparing students in the area of computer science is the development of student's programming skills. Despite a heroic academic effort and many different theoretical researches which deal with the question of how to develop these skills, a substantial minority of students fails introductory programming course. We observed this fact at our university and noticed it's similar at other universities. Research concerning students' programming capabilities evaluation with regard to the input Dehnadi's test focused on programming aptitude has been done at the Faculty of Informatics and Management and at the Faculty of Science, University of Hradec Králové, since 2011. Within the research most often students' mistakes have been deeply examined and study materials has been innovated. This paper is aimed to study materials innovation with regards to the main student mistakes and student learning styles. It could serve to pedagogues as an inspiration to their teaching in the mentioned area.

Keywords Development of algorithmic thinking, student learning styles, algorithmic thinking evaluation.

I. INTRODUCTION

DEHNADI et al. developed a programming aptitude test which predicts whether a person can become a programmer. The test does not require any prior programming experience which is an important property that allows student pre-screening and teaching adjustments before a programming course begins. We have pursued a research project focused on programming capabilities evaluation using Dehnadi's approach at the Faculty of Informatics and Management and at the Faculty of Science, University of Hradec Králové since 2011. During the research we deeply examined the most common mistakes students make and innovated the study materials according to the research results.

Firstly, we briefly introduce the subject Algorithms and Data Structures which is our introductory programming course at both faculties. Secondly, we present the main results of evaluating student programming capabilities using Dehnadi's test. We follow with a description of the most common learning styles that are preferred by the students of our university. We conclude this paper with a summary of innovations of our study materials with regards to the research results and preferred learning styles.

II. ALGORITHMS AND DATA STRUCTURES

The Algorithms and Data Structures (ALGDS) course is aimed at basic algorithm construction skills. It is focused on novices in the area of programming.

We use the brick-box approach to explain basic algorithmic constructs where only several base elements are available from which children (students) are able to create incredible buildings (programs). We use a Czech pseudo-language which is a modified Pascal programming language with basic commands. (Remark: The Pascal programming language was created by Nicklaus Wirth specifically for educational purposes, cf. [1])

A. Lectures and seminars

We explain all basic algorithmic structures in the lectures. We start with those that use single variables. We properly explain the idea of each presented algorithm and usually demonstrate it with the help of students (e.g. each student tells a value and the action of the algorithm is introduced), as well as illustrate the progress of each algorithm in a step-by-step procedure. After the thorough practising of basic algorithmic structures on single variable problems we present one-dimensional and two-dimensional arrays. Using these data structures we also repeat all previous algorithms with a careful attention to manipulating array indices.

The students apply their acquired knowledge during seminars to a variety of tasks. They work in groups of two or three and each group is responsible for solving one of the given tasks. The students are given some time to prepare their solutions on a piece of paper. Then each task is illustrated and presented by two students (or three, depending on the number of groups, each group responsible for the task-solution deposes one student) on the blackboard. All students discuss and compare the given solutions. This approach is beneficial both for the students who are forced to try and find more solutions to the given task and for the teacher who has an opportunity to open a discussion should there be a problem in the presented solutions. For more information see [2].

B. Study materials

The whole area explained within the subject ALGDS is introduced in the textbook [3], where more than 150 problem assignments, questions and exercises are presented. The accuracy of a solution can be verified with the help of the *Algorithms* program which is attached together with solutions

of all the textbook's given tasks on a CD.

Students can view the electronic version of the textbook as well as download the *Algorithms* program and other study materials corresponding to a lecture (additional lecture notes, problem statements of tasks solved in seminars, presentations and animations used at the appropriate lecture as a complement of the explained matter) in the virtual study environment used at the university.

There is also a detailed plan of lectures available as well as credit and exam conditions and samples of credit and exam tests.

III. DEHNADI'S TEST AND RESEARCH RESULTS

Introductory programming courses at universities often face the problem of bimodal distribution of the applicants score. There seems to be a clear division between people who find programming easy and those who find even basic programming tasks difficult. Dehnadi and Bornat observed that the level of mental model consistency that learners apply when faced with a novel, seemingly nonsensical problem could be a good predictor of programming capabilities (see [4]). Saeed Dehnadi designed a test focused on discovering future programmers before they enter their first programming class [5].

A. Dehnadi's test

Dehnadi's test [6] is based on assessing mental model consistency in the assignment operation. The test consists of twelve similar questions. Each question gives a sample C-like program, declaring two or three variables and executing up to three variable-to-variable assignment instructions. The test is evaluated according to an answer sheet and a mark sheet which assign mental models to each answer and combine them into a total score. There are 11 mental models described in [6]. Except for the classical assignment operation model known from C or Java, Dehnadi identified 10 other models that were used by the applicants. The common ones include left to right assignment (as opposed to the standard right to left), comparison or assignment without carry (the assignment effects are not carried to the next line). If a person is able to choose any of these 11 models and use it consistently through the whole test, he or she is considered a good candidate to become a programmer. The consistency threshold was set at 8 consistent answers out of 12. Dehnadi and Bornat [4] report that 44 % of the applicants showed consistent models, 39 % used several models inconsistently and 8 % have refused to fill the answers. The remaining 9 % of students are not discussed in the paper." (see [7])

B. Research sample

In the academic year 2011/2012 we decided, for the first time, to use Dehnadi's test at the Faculty of Informatics and Management to test our freshmen at the beginning of the first programming ALGDS course. We hoped that Dehnadi's test could serve as an orientation resource both for students and teachers. Students who belong to the inconsistent group

should devote a bigger attention to the subject; they should study regularly, discuss their solutions with the teacher or experienced students and practice their programming skills (preferably every day). Teachers would be able to use the Dehnadi's test results to divide students into appropriate groups when explaining and practicing algorithm design. During the academic years 2011/2012, 2012/2013 and 2013/2014 data of 343 students who both filled Dehnadi's test and participated in the introductory programming course were analysed.

In the academic year 2012/2013 we started to use Dehnadi's test also at the Faculty of Science to test our first-year students, future teachers studying the Informatics specialization. We analysed 59 student results in the academic years 2012/2013 and 2013/2014.

Because of the fact that there are students coming to the university who already have a prior programming experience before attending the first programming course we divide the students to two groups: *students with no prior programming experience* and *students with prior programming experience*.

C. Results and discussion

The analysis using Wilcoxon rank sum test and additionally, a Spearman's rank correlation test was performed.

Contrary to the claims in [4], results of our research have shown that the Dehnadi's test *cannot* predict success in the ALGDS course for *students with no prior programming experience* (cf. [8]).

Nevertheless using the test *we are able to predict success in ALGDS subject of students with prior programming experience*. Supposing that about half students attending ALGDS subject at both faculties have prior programming experience, the Dehnadi's test can serve as a useful information for students not achieving full score to devote bigger care to the subject. [9]

Based on many years' experience we are sure that the existence of students who successfully passed Dehnadi's test but did not succeed in ALGDS can be explained as follows: Students with prior programming experience usually think that their previous programming skills are enough to pass the ALGDS subject and do not devote the necessary attention to the subject matter.

On the other hand students who do not succeed in the Dehnadi's test are able to pass ALGDS subject thanks to a systematic development of algorithmic thinking. We found out that devoting enough time explaining mutual relationships among solved problems and carefully, together with students, recognizing the differences among almost the same algorithms is very helpful for students.

IV. LEARNING STYLES

Last several years the Index of Learning Styles (ILS shortly) has been used at our university both at the Faculty of Informatics and Management and at the Faculty of Science. ILS was developed by Richard M. Felder and Barbara A.

Soloman and it may be used at no cost for non-commercial purposes by individuals who wish to determine their own learning style profile and by educators who wish to use it for teaching, advising, or research. ILS is an on-line instrument used to assess preferences on four dimensions (*active/reflective*, *sensing/intuitive*, *visual/verbal*, and *sequential/global*). It is based on a model developed by Dr. Richard M. Felder in collaboration with Dr. Linda K. Silverman, generally referred to as the Felder-Silverman learning style model. For more information see [10] and [11].

The main results concerning student learning styles gained from three previous years 2012–2014 at the above mentioned faculties can be summarized very briefly as follows (see [12], [13]): the *sensing* and *visual* learning style dominates student preferences (more than 80% students).

Let us quote [14]:

Sensing / intuitive learners

Sensing learners tend to like learning facts, intuitive learners often prefer discovering possibilities and relationships.

Sensors often like solving problems by well-established methods and dislike complications and surprises; intuitors like innovation and dislike repetition. Sensors are more likely than intuitors to resent being tested on material that has not been explicitly covered in class.

Sensors tend to be patient with details and good at memorizing facts and doing hands-on (laboratory) work; intuitors may be better at grasping new concepts and are often more comfortable than sensors with abstractions and mathematical formulations.

Sensors tend to be more practical and careful than intuitors; intuitors tend to work faster and to be more innovative than sensors.

Sensors don't like courses that have no apparent connection to the real world; intuitors don't like "plug-and-chug" courses that involve a lot of memorization and routine calculations.

Everybody is sensing sometimes and intuitive sometimes. Your preference for one or the other may be strong, moderate, or mild. To be effective as a learner and problem solver, you need to be able to function both ways. If you overemphasize intuition, you may miss important details or make careless mistakes in calculations or hands-on work; if you overemphasize sensing, you may rely too much on memorization and familiar methods and not concentrate enough on understanding and innovative thinking.

Visual / verbal learners

Visual learners remember best what they see--pictures, diagrams, flow charts, time lines, films, and demonstrations. Verbal learners get more out of words--written and spoken explanations. Everyone learns more when information is presented both visually and verbally.

In most college classes very little visual information is presented: students mainly listen to lectures and read material written on chalkboards and in textbooks and handouts. Unfortunately, most people are visual learners which means

that most students do not get nearly as much as they would if more visual presentation were used in class. Good learners are capable of processing information presented either visually or verbally.

V. INNOVATIVE STUDY MATERIAL FOR THE ALGDS SUBJECT

We mentioned in section II that students can download the textbook [3], the *Algorithms* program and lecture notes from the virtual study environment.

We used the results gained from analysing student learning styles to order the exercises for a given textbook topic from *easier* to *more difficult* ones. We carefully discuss *mutual relations among algorithms* at the lectures and seminars.

Until the academic year 2013/14 we had first explained the one-dimensional data structure. After the thorough practising we had proceeded to two-dimensional array problems. In 2015 we decided to change this process. We explain each typical algorithmic structure for both one-dimensional and two-dimensional array. We found out that the students are able to better comprehend the array index manipulation this way and thus better understand the whole subject matter. Compared to the previous years *the fear of working with two-dimensional array indices almost disappeared*.

To support more and more visual learners we have prepared another significant innovation. The texts describing the typical array manipulation techniques were complemented by *colours* emphasizing structures used in the given algorithm. Let us introduce this concept on the typical algorithmic structure *Inserting a new array element*.

Inserting a new array element

Let us introduce a typical algorithmic structure used in algorithms for adding a new value x after the k -th term of a numerical sequence (a_1, a_2, \dots, a_n) , $1 \leq k \leq n$, stored in an array a .

```
read(x);
read(k);
i := n;
while i ≥ k + 1 do
begin
  a[i + 1] := a[i];
  i := i - 1;
end;
a[k + 1] := x;
n := n + 1;
```

Example

Let us create an algorithm which adds -1 after the **last minimum element** of a numerical sequence (a_1, a_2, \dots, a_n) , $1 \leq k \leq n$, and **writes out the resulting sequence**.

```
begin
  read(n);
  for i := 1 to n do
    read(a[i]);
```

```

min := a[1];
indexMin := 1;
for i := 2 to n do
  if a[i] ≤ min then
    begin
      min := a[i];
      indexMin := i;
    end;

```

```

i := n;
while i ≥ indexMin + 1 do
  begin
    a[i + 1] := a[i];
    i := i - 1;
  end;
a[indexMin + 1] := -1;
n := n + 1;
for i := 1 to n do
  write(a[i]);
end.

```

VI. CONCLUSION

We presented a detailed list of the improvements to our introductory programming course at University of Hradec Králové that stemmed from our four-year research effort into programming aptitude testing and learning style analysis.

We showed that while early programming aptitude tests cannot be used blindly on all students, good discriminative results can still be achieved for students with a prior programming experience. We can focus on the students with prior programming experience who however fail at the simple programming aptitude test and improve their success rates in the course.

The research of learning styles led to a technically simple but an extremely helpful extension of the textbook that uses colour coding to match sentences from the exercise with corresponding lines of its algorithmic solution.

Results achieved from the credit and exam tests show that our improved pedagogical approach used in the first programming course, Algorithms and Data Structures, is successful.

ACKNOWLEDGMENT

This research has been supported by Specific research project of the University of Hradec Kralove, Faculty of Science No. 2113.

REFERENCES

- [1] Wirth, N. (1975) *Algorithms + Data Structures = Programs*, Prentice-Hall, New Jersey.
- [2] Milková, E. (2015) Multimedia Application for Educational Purposes: Development of Algorithmic Thinking. *Applied Computing and Informatics* Vol. 11, Issue 1, 76–88. doi: 10.1016/j.aci.2014.05.001
- [3] Milková, E. et al. (2010) *Algoritmy: typové konstrukce a příklady*. Hradec Králové: Gaudeamus.
- [4] Dehnadi, S. and Bornat, R. (2006) The camel has two humps (working title), Middlesex University, UK
- [5] Dehnadi, S., Bornat, R. and Adams, R. (2009) Meta-analysis of the effect of consistency on success in early learning of programming. *Psychology Programming Interested Group (PPIG) Annual workshop*.
- [6] Dehnadi, S. (2006) Testing programming aptitude, *Proceedings of the 18th Annual Workshop of the Psychology of Programming Interest Group*, pp. 22–37.
- [7] Milková, E., Petránek, K. and Janečka, P. (2012) Programming capabilities evaluation, *Proceedings of the 9th International Conference on Efficiency and Responsibility in Education*, Czech University of Life Sciences Prague, pp. 310–318.
- [8] Bennedsen, J. and Caspersen, M. E. (2007) Failure rates in introductory programming, *ACM SIGCSE Bulletin*, vol. 39, no. 2, pp. 32–36
- [9] Milková, E., Kořínek O. (2013) Students' programming capabilities evaluation. *Proceedings of the 10th International Conference on Efficiency and Responsibility in Education*, Czech University of Life Sciences Prague, pp. 434–440.
- [10] http://www4.ncsu.edu/unity/lockers/users/f/felder/public/Learning_Style_s.html
- [11] <http://www4.ncsu.edu/unity/lockers/users/f/felder/public/ILSpa.html>
- [12] El-Hmoudova, D. (2015) Assessment of Individual Learning Style Preferences with Respect to the Key Language Competences, *Procedia - Social and Behavioral Sciences*, vol. 171, 16 January 2015, 40-48. doi:10.1016/j.sbspro.2015.01.103
- [13] ElHmoudová, D. (2013) Assessment of individual learning style preferences in blackboard environment, *Proceedings of the 10th International Conference on Efficiency and Responsibility in Education*, Czech University of Life Sciences Prague.
- [14] Felder, R. M., Silverman L. K. (1988) Learning and Teaching Styles in Engineering Education, *Engr. Education*, 78(7), pp. 674–681.

Interactive Teaching Tools for Visualizing Geometrical 3D Objects Using Pseudo Holographic Images

M. Ciobanu, A. Ploscar, I. Dascal, I. Virag and A. Naaji

Abstract— Starting from the fact that the mathematical abstract notions are often hardly understood by students and influence their learning outcomes, we created and implemented a set of interactive teaching tool packages (ITTPs) consisting of theoretical modules and applications in the framework of a Hungarian - Romanian cross-border project. One of the main ideas was to develop all the applications as open source software. The Romanian team focused on basic elements of linear algebra and analytic geometry with some relevant topics, using an autostereoscopic display for image visualization. In this paper we will present the applications for viewing geometric objects, namely conics and intersections of straight lines, planes and spheres, by pseudo holographic images. We extended the concept to work inside a web browser in order to eliminate the need to install any other software. All the ITTPs are available on the partner universities e-learning platforms and project web page.

Keywords— autostereoscopic display, teaching tools, pseudo holographic images

I. INTRODUCTION

In the last years there has been an increasing development of online courses, most of them with interactive components. By 2011 the massive open online courses (MOOC) movement was materialized in the form of multiple platforms offering open courses to large numbers of students, using technologically enhanced resources and activities, such as virtual laboratories, simulations and cognitive tutors. Hundreds of courses are available all over the world on MOOC platforms such as Coursera, edX, or Udacity. Rhoades et al.[1]

M. Ciobanu is with Department of Computer Science, “Vasile Goldis” Western University, 310025 Arad, Romania (e-mail: cmmciobanu@gmail.com).

A. Ploscar is with Department of Computer Science, “Vasile Goldis” Western University, 310025 Arad, Romania (e-mail: adina_ploscar@yahoo.com).

I. Dascal is with Teoretical Highschool Pancota, 315600 Pancota, Arad, Romania (e-mail: nelu_dascal1@yahoo.com).

I. Virag is with Automation and Computer Science Faculty, Polytechnica University of Timisoara, 300006 Timisoara, Romania. (e-mail: ioan.virag@aut.upt.ro).

A. Naaji is with Department of Computer Science, “Vasile Goldis” Western University, 310025 Arad, Romania (+40-257-285813; e-mail: anaaji@uvvg.ro).

considers that the open educational resources (OER) movement have helped increase the availability of a great number of courses and educational materials at a global level, holding the potential to „revolutionize how higher education practitioners, scholars, and policymakers think about and define democratic forms of access”.

In this context, considering the potential and the benefits of e-learning platforms, we created and implemented some courses for mathematical disciplines, adapted to online education. In order to facilitate understanding and learning abstract mathematical notions, interactive teaching tools packages (ITTPs) have been developed, containing both theoretical modules and applications. Some ITTPs are related to mathematical analysis, probability, numerical calculus, computational algebra, the corresponding applications being made using Maxima. ITTPs referring to concepts of linear algebra and analytical geometry contain applications created in Java (for Android), Java 3D and JavaScript. All of the developed applications are open-source.

These results were obtained in the framework of a Hungarian - Romanian cross-border project (see Acknowledgement). The Romanian team has developed three sets of interactive courses which focus on basic elements of linear algebra and analytic geometry. Relevant topics are: linear spaces, straight lines and planes, conics and quadrics.

The interactive applications conceived for these topics facilitate the understanding of abstract notions, following the perception gained by the student from direct interaction with the three-dimensional geometric object. The applications created within the project provide the possibility to regard this object from different angles.

An autostereoscopic display was used for visualizing 3D objects as holographic images, as well as the outcome of parameter changes inside mathematical formulas. For each application we created a version which runs on an ordinary computer and another one adapted for the holographic device. Thus, this device is used for the first time for educational purposes.

In developing these sets of courses, the aim was to give a detailed overview of the concepts, with examples that make the content more accessible. The purpose of these ITTPs was to provide students with a new tool in order to help them understand and learn mathematical notions.

II. SOME TECHNICAL ISSUES

The use of holographic-like images is a cutting-edge technological approach, with applications in different fields such as medicine, industry, military, advertising, entertainment, computer game development etc. We used this technology to improve the quality of education by developing new visual-interactive methods for teaching mathematics.

In their research, Wolfgang Schlaak from the Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute [2] built an application that allowed physicians to study the patients' CT scans by visualizing them in 3D, with the possibility of rotating the image by means of hand gestures (called non-contact image control). The display they used was an autostereoscopic multiview display, which involved tracking the user's eyes by means of a set of cameras mounted on top of the display and adjusting the image accordingly, to produce a sensation of depth.

In developing our applications we used a HoloVision aerial image display by Provision [3]. This new type of autostereoscopic display does not require any cameras for tracking the viewers' eyes, being based on Pepper's Ghost technique [4], and allows a larger number of persons to visualize the projected 3D models without the use of special glasses.

Before our project implementation, to our knowledge, the HoloVision display was only used for marketing purposes; we are the first scientific team that used the aerial image display for educational purposes.

The main difference between the two technologies is that the monitor we used does not have the 3D image attached to the display, but rather exhibits similarity to an actual hologram. The monitor uses a special mirror in order to focus and display the pseudo 3D holographic image.

The HoloVision holographic device we used is able to display a 30 cm image focused at approximately 90 cm in front of the monitor, viewable at a 60-degree angle from the center and doesn't require the user to wear any kind of special glasses [5].

III. THE APPLICATIONS DEVELOPED FOR INTERACTIVE TEACHING TOOL PACKAGES

In this paper we present some of the applications related to analytical geometry [6], [7], namely straight lines, planes, spheres and conics, which use autostereoscopic display. The applications were developed in two versions, one for individual study on a 2D monitor and one for presenting the topics to a group of students using the 3D display, as the viewing angle and distance allow this. It was necessary to create two types of interfaces because the HoloVision display is optimized to render 3D models focusing the image outside of the display area, providing the illusion that the image floats in midair in front of the monitor. The 3D image cannot be rendered in the same way on a 2D monitor, but the 3D model keeps its details on the HoloVision display even if it is scaled. In Fig. 1 we present both versions of an application.

The applications were developed in Java 3D API or JavaScript.

The Java 3D API is an application-programming interface used for creating portable 3D applications and applets. It provides a library for creating and manipulating 3D models and for constructing the structures used in rendering the scene. The scene is constructed in the form of graph whose nodes are objects that will be included in the application. The graph associated to the scene is structured in the form of a tree whose root is a virtual world. Developers can easily describe very large virtual worlds using this approach [8].

JavaScript is a programming language used to make interactive web pages [9]. It can run inside any up-to-date browser, and also allows extending the applications for mobile devices.



Fig. 1 interfaces of the applications developed for holographic display and laptop

A. The applications developed for representing straight lines, planes and intersections between them

The ITTP Straight lines and planes in 3-dimensional Euclidian space presents the characterization of lines and planes by vector and Cartesian equations; formulas are given to calculate the distance between two points, the distance from a point to a plane, the distance from a point to a line, the distance between two non-coplanar lines, as well as to calculate the angle between two lines, between two planes, and between a line and a plane.

Fig. 2 shows an application which allows generating the image of a straight line by introducing the coordinates of a point and a direction vector, and the image of a plane, respectively, by introducing the coefficients of the Cartesian equation.

In order to obtain two planes or lines, the requested data are introduced successively, and the application checks whether the objects intersect each other, in which case it displays the angle between them in radians and the coordinates of the point or line of intersection.

Another application generates the common perpendicular line between two non-coplanar straight lines and thus allows calculating the distance between them.

The applications were created using Java 3D API.

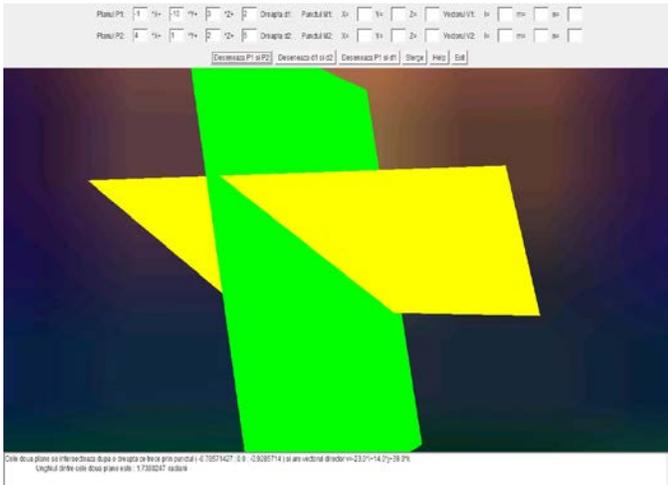


Fig. 2 interface of the application developed for visualizing straight lines and planes

B. The applications developed for representing spheres and the circle in 3-dimensional space

In order to generate the image of a sphere, our application entails entering the coordinates of the center and radius of the sphere. The program calculates and displays the power of a point with respect to a sphere (Fig. 1). Data can be given successively to represent two, three or four spheres. Correspondingly, the program generates the radical plane, the radical axis and the radical center, and their equations or coordinates respectively are displayed.

Another application presents the intersection between a sphere and a straight line as being the solutions to the system consisting of their equations. In this case, the system is reduced to a second degree equation which has the most two real solution, and thus we have the following three situations: the equation has two distinct real roots – the line is secant, i.e. it intersects the sphere in two distinct points; the equation has a double root – the line is tangent to the sphere; the equation does not have any real roots – the line does not intersect the sphere.

In case of intersection between a sphere and a plane, a circle is obtained in the 3-dimensional space. The corresponding application, whose interface is shown in Fig. 3, allows viewing all possible cases, namely: when the distance from the center of the sphere to the plane is larger than the radius, the plane does not intersect the sphere; when the distance from the center of the sphere to the plane is equal to the radius, the plane is tangent to the sphere; when the distance from the center of the sphere to the plane is smaller than the radius, the plane is secant. In this last case, the application generates the outline of the circle and displays the coordinates of the center and radius of the circle.

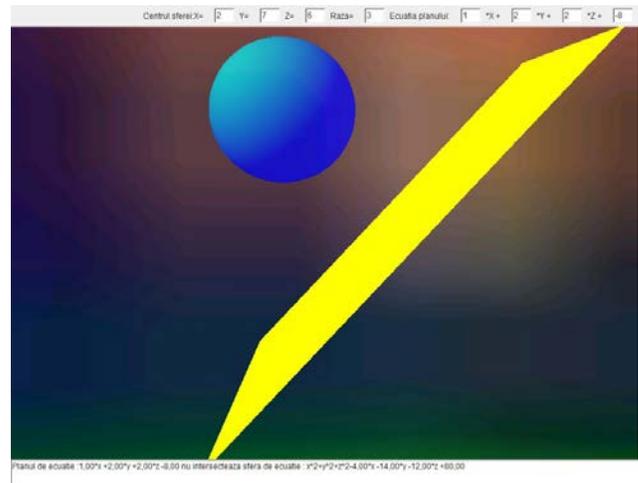


Fig. 3 interface developed for visualizing position of a sphere to a plane

The circle can also be obtained as an intersection of two spheres. The application let us analyze all possible situations regarding the distance between the centers of the two spheres.

These applications were also made using Java 3D API. The images can be rotated, thus allowing visualization of 3D model from different angles.

C. The applications developed for representing conics

In the ITTP presenting conics, canonical equations are given for non-degenerate and degenerate conics, respectively, and the reduction of a conic's equation to its canonical forms is presented. Orthogonal invariants are defined too, in order to establish the nature and genus of a conic.

For the graphic representation of conics given by general equations, an application allows determining the canonical equation of the conic, the center or vertex of the conic and the axes or axis of symmetry. To this end, orthogonal invariants are calculated and, depending on their values, several cases are possible. Examples are presented for the representation of conics with center – ellipse and hyperbola, intersecting lines, and with no center (unique at finite distance) - parabola, parallel lines, together with complete solutions.

Using JavaScript, we developed applications for representations of conics as intersections of a cone surface and a plane. It is possible to rotate the whole 3D model for viewing the intersection of the plane with the cones from different angles. The user interface also allows the movement of the plane on the Ox and Oy axes, or its rotation around the Ox axis. Conic sections such as the circle, ellipse, parabola or hyperbola can be viewed in real time by adjusting the cursors (Fig. 4 and Fig. 5).

The source code for the applications is based on standards like HTML5, CSS, the jQuery API [10] for the user interface, the Raphael library [11] for working with vector graphics and the Conics3D JavaScript extension written by Lodewijk Bogaards [12] for the visualization of conic sections.

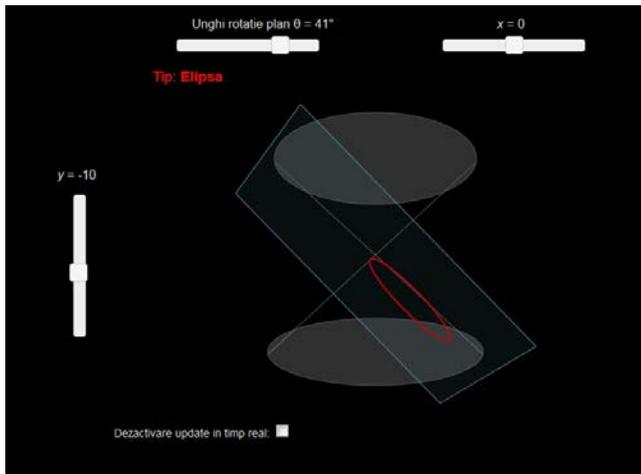


Fig. 4 interface for visualizing a conic section (an ellipse) by movement of the plane on the Ox and Oy axes

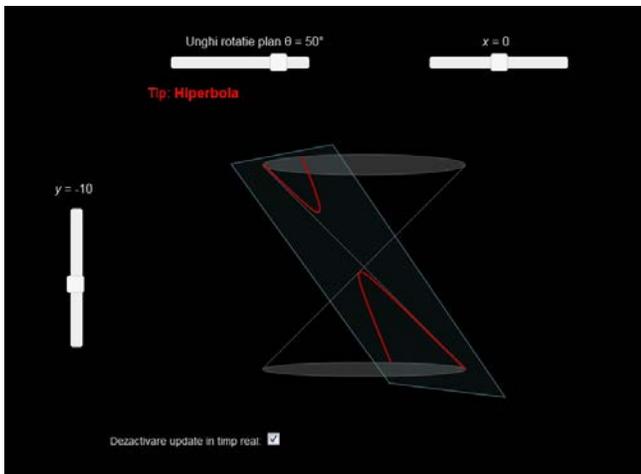


Fig. 5 interface for visualizing a hyperbola and the rotation of 3D model around Ox axis

IV. CONCLUSIONS

Considering the rapid development of e-learning worldwide, within our project, we created some online courses for mathematical disciplines, designed as interactive teaching tools packages (ITTPs). We attempted to make the presentation of the topics as student-friendly as possible, so all formulas or theoretical solution methods highlighted were backed by examples.

For the applications related to analytical geometry, the use of autostereoscopic display provides a new approach both from the technical point of view and teaching methods. The interactive applications conceived for the topics of these ITTPs facilitate the understanding and learning the mathematical subjects.

As future work we intend to develop new applications for visualizing 3D geometrical objects as pseudo holographic imagines, using the autostereoscopic display, and also to extend our research in the medical field.

ACKNOWLEDGMENT

The presented results were financially supported by the Hungary-Romania Cross-Border Co-operation Programme 2007-2013, under the grant no. HURO/1001/040/2.3.1., entitled: Development of Open Source Software and Interactive Teaching Tools for Mathematics – SOFTMAT (<http://softmat.uvvg.ro>). The project was implemented by College of Nyiregyhaza, Hungary as lead partner and “Vasile Goldis” Western University of Arad, Romania.

REFERENCES

- [1] R.A. Rhoads, J. Berdan and B. Toven-Lindsey. The Open Courseware Movement in Higher Education: Unmasking, Power and Raising Questions about the Movement's Democratic Potential, *Educational Theory*, vol. 63, no. 1, pp. 87-109, 2013
- [2] ***, *Fraunhofer Institute of Research* http://www.fraunhofer.de/archiv/pi-en-2004-2008/EN/press/pi/2007/11/ResearchNews112007_Topic3.html
- [3] ***, *Provision Company* webpage (<http://www.provision.tv/>)
- [4] J. Nickell, *Secrets of the Sideshows*, University Press of Kentucky, 2005
- [5] M. Ciobanu, A. Naaji, I. Virag and I. Dascal. Interactive Applications for Studying Mathematics with the use of an Autostereoscopic Display, in *Soft Computing Applications, series Advances in Intelligent and Soft Computing*, Springer Berlin, Heidelberg, to be published
- [6] D.Andrica and L. Topan. *Analytic Geometry*, Cluj University Press, Cluj-Napoca, 2004
- [7] Gh. Atanasiu, Gh. Munteanu and M. Postolache. *Algebra liniara, geometrie analitica si diferentiala. Ecuatii diferentiale*, Editura Fair Partners, Bucuresti, 2003
- [8] L. Ammeraal, and K. Zhang, *Computer Graphics for Java Programmers*, John Wiley & Sons Ltd, 2007
- [9] A. Osmani (2014). *Learning JavaScript Design Patterns*, [Online]. Available: <http://addyosmani.com/resources/essentialjsdesignpatterns/book/>
- [10] ***, *jQuery API Documentation*, <http://api.jquery.com/>
- [11] ***, *The Raphaël JavaScript library home page*, <http://dmitrybaranovskiy.github.io/raphael/>
- [12] L. Bogaards, *Conic Sections app*, <http://www.mrhobo.nl/>

Implications of Domain-driven Design in Complex Software Value Estimation and Maintenance using DSL Platform

Nikola Vlahovic

Abstract—The introduction of the domain-driven design (DDD) as an alternative approach to software development had the promise of achieving several benefits in the process of creating complex domain-specific business applications. Due to the focus of this approach to the core of the application functionality, improved collaboration with domain experts and conceptual modeling benefits, it has attracted a reasonable amount of attention from the programming community in the past decade. Aforementioned benefits have also been able to create unique set of programming environments and languages that also move the boundaries of efficiency of code execution and application maintenance.

In this paper we will present and analyze one such tool, namely, DSL platform. DSL platform is a service that allows for the design, creation and maintenance of business applications. The goal of this paper is to analyze the implications of using the DDD through the DSL platform on several important aspects of software management. Primarily we will focus on the estimation of complex software system value and software refactoring and maintenance effort based on the models proposed by Groot et al.

We will show that for complex software systems consisting of a number of different components, programming paradigms and database systems can highly benefit from this approach. Some of the most important benefits pertain to lowering of the cost of software maintenance and transcending the properties of reliable business applications and databases developed using legacy systems to current systems using the underlying domain model.

Keywords— Software development, Software value, Software maintenance, Domain-driven design, Software engineering, Software refactoring, Legacy systems.

I. INTRODUCTION

THERE is a reasonably limited number of papers in current scientific literature pertaining to different particularities of software development management and practices, such as software pricing model practices or adoption of novel software development approaches. Only in recent years overviews of some of these aspects of software management have been studied and new models have been proposed. At the same time practitioners are developing and presenting new frameworks and technologies as well as new approaches to software development altogether. Only a limited number of these

developments enter the mainstream adoption by software or even non-software companies.

One such phenomenon is software development approach called domain driven design. This approach tries to offer solutions to bridging the gap between business experts and software experts that is main drawback in traditional approaches that decreases the success rate of many software projects. Agile methodologies are more successful in coping with this gap for reasonably limited and small-scale software systems. When it comes to complex business systems only approaches with traditional core principles are available, mostly with increased inefficiency and additional development and maintenance costs.

Domain driven design is therefore dedicated in improving the development and maintenance efficiency in complex business systems. While offering great benefits for this type of software systems and additional improvements in various aspects of software management, it still faces significant obstacles to adoption.

In this paper we will explain and present the main concepts of domain driven design as an adequate software development approach for complex business systems. Implementing these benefits will be given through description of one particular implementation of the approach, a software development tool called DSL Platform. Benefits can be critically assessed through different software management issues, and in this paper we will concentrate on estimation of software asset value and maintenance of these assets.

Goal of this paper is to critically investigate possible benefits of adopting domain driven design in software management, with particular emphasis on maintenance during production phase of software assets. Inevitably these considerations will reflect on the value of software asset, so a validated approach to estimation of software assets is called for. Here we will build on a proposed model of software valuation proposed by [2].

The structure of this paper is as follows: In Section II we will describe the main features of domain driven design, its advantages and disadvantages as well as the implementation of its concepts in a tool called DSL Platform. In Section III we will take a closer look at some of the most important software management issues that can be affected by the domain driven design approach, such as the software maintenance issues,

N. Vlahovic is the associate professor at the Informatics Department of the Faculty of Economics and Business, University of Zagreb in Croatia. Trg. J.F. Kennedyja 6, 10000 Zagreb, Croatia (phone: +385-1-238 3220; fax: +385-1-233 5633; e-mail: nvlahovic@efzg.hr).

software risk managements and software value assessments approaches. In Section IV we will discuss the possible impacts of applying domain driven design within the software development process for complex coupled heterogeneous systems throughout the software process life cycle and extrapolate the benefits and issues that the management should be aware when considering introduction of domain driven design. Finally in Section V. conclusions will be given with a few guidelines outlining the main advances DDD and DSL Platform can provide for companies using complex heterogeneous software systems.

II. DOMAIN DRIVEN DESIGN AND DSL PLATFORM

In this section we discuss the domain driven design as a type of software development approach, position this approach in a wider context and based on our analysis describe a tool that implements these features in most consistent manner.

Domain driven design (DDD) is a software development approach that rather than analytically organize the software development effort and use conceptual, modeling, programming and implementation tools, it tries to make a complete model of the problem domain moving the focus of the development effort away from tools, techniques and methodologies used.

In the most general terms software development approaches can be divided into two diametrically contrasted classes and one intermediary class that draws on some of the concepts from either of the two main classes [1]:

- 1) Class of structured approaches. This is a group of software development methodologies that are based on a process that recognizes distinct phases of the software development process. These phases usually align with particular stages of the software development life cycle (SDLC). Depending on the particular methodology each phase can be associated with a stage in SDLC either, planning, creating, testing or deploying of the software system. Some methodologies can have several phases associated with one stage of the SDLC, and others can have one phase spanning over or overlapping with two stages of the SDLC. The main characteristic of methodologies in this group is that each phase needs to be completed with some final result, a software artifact, before next phase of the process can begin. Some of the most common methodologies that belong to this group are waterfall software development model, prototyping, incremental development, iterative incremental development, Boehm's spiral model, etc. but also object oriented approaches.
- 2) Class of behavioral approaches. This group of methodologies relies on the soft systems approach that takes a more relaxed definition of development process. Behavioral approaches take a holistic view of the organizational systems and social nature of software systems (both in development and deployment stages).

This is why these methodologies promote participation of system users and customers during the creation phases of the system. Also the development process may return to earlier phases as required by the current perspective of the software system and even different development activities may overlap. Along with soft systems approach we can find characteristics of the behavioral approach in agent based software engineering [3], [4] as well as in the behavior-driven design [5].

- 3) Intermediary and transitional approaches. This class of approaches to software development shares some of the characteristics with the structured approaches and some of the characteristics with the behavioral approaches. These methodologies represent the synthesis of traditional rigid structure and softer humanist elements of the behavioral approaches. Agile methodologies represent the most typical example of a transitional approach due to their strive to capture the human aspects of organization for all stakeholders involved, especially during the analysis and planning stages, while still retaining structure in design and implementations stages [6], [1].

Domain driven design (DDD) as a somewhat recent novel software development approach tries to change the traditional focus from the project methodologies and tools towards the core of the problem at hand. DDD goes even beyond a particular technology or methodology, or even a framework. It is a way of thinking and a set of priorities aimed at accelerating software projects that have to deal with complicated domains [7]. As such it is very close to behavioral approaches, but as it strongly relies on hierarchies of priorities and concepts typical for structured approaches, it can be regarded as a transitional approach to software development. Still, unlike agile methodologies that are focused on a limited, small to medium sized software projects, DDD is primarily concerned with complex and coupled software systems. As it is platform-independent it is an encompassing approach to highly coupled systems that use different, even inconsistent, technologies and platforms as well as development methodologies or practices.

In order to understand how DDD can connect all of the varieties of concepts into a consistent and unified one we will take a look at how previous methodologies and frameworks represent software projects. Most of them treat a software project as an entity that has to be described using a number of different perspectives. Since there are a lot of different stakeholders involved in the development of any software project, a variety of perspectives is used to promote better communication and understanding between stakeholders. In practice Unified Modelling Language (UML) is mostly used for static and dynamic representation of these perspectives. UML covers all of the relevant views of the software system, its surroundings and dependencies using three groups of dedicated diagrams, structure diagrams, behavioral diagrams and interaction diagrams [8]. Inevitably, different perspectives

may not be entirely compatible and this may present a challenge for the development team in continuation with the development of the project.

Unlike UML that takes on a number of perspectives of the model, DDD tries to describe the model by describing its

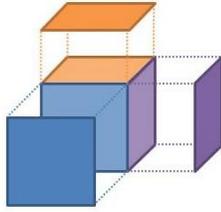


Fig. 1 model and perspectives of the model

domain as a whole and complete model (Figure 1). In this way, model itself represents the system being developed. Consequentially, programming code is the representation of the model. Inappropriate, platform-dependent technical programming code would cause lock-out effect for diversity of technologies, platforms, methodologies as well as a number of stakeholders, especially business experts with no programming skills.

In order not avoid these lock-out effects specific requirements are expected from the team communication facilities. Firstly a domain specific language (DSL) is required to describe the model of the software project, and secondly a ubiquitous language for team communication should be used and evolved during the development of the project. Consistent communication between business domain experts and developers expressing their views of the system in terms of model concepts will evolve in a ubiquitous language. The team understanding of software artefacts will express itself in the source code of the system as it represents the model of the system (through DSL). Any change in the model will change the model and these changes are clearly visible to all of the project participants, both business experts and developers [9]. DDD is an ongoing process of expressing ubiquitous domain language in code [10].

Implementing key features of the DDD using object oriented design can be used to create a unified platform for development and evolution of complex software systems. One such tool is DSL Platform which we will describe in the rest of this Section.

DSL Platform is a service that helps in designing, building and maintaining business applications. It allows for the automation of business application development process. The platform uses the specific business model as input and outputs finished components for corresponding business software system. Since DSL platform draws on the strengths of the DDD approach, business model is described in understandable language for both business experts and development team while this description is also a formal specification of the system (Figure 2).

Once declarative specification is defined, any of the supported compilers can use this specification to build code or



Fig. 2 DSL Platform concept

maintain databases. True value of DDD approach becomes apparent during the maintenance and evolution of the system. Any changes made to the business model are automatically translated by the platform into Client code or Databases (as shown in Figure 2). This functionality alleviates programmers' efforts and moves focus of their work to specific functionalities and user experience rather than code optimization, refactoring or similar technical tasks. Similarly the maintenance or even migration of data to the underlying database system is also highly automated.

Two main challenges that can be effectively solved using DSL Platform and underlying DDD approach is the elimination of miscommunication between clients and contractors or even among developers within developer teams. The other is the elimination of non-creative and repetitive work done by developers by automating repetitive tasks of the development process.

A. Tackling miscommunication

In each software project there is a number of different stakeholders that need to communicate their views, ideas and concepts between themselves. Due to different backgrounds (business backgrounds or engineering backgrounds) as well as different perspectives of the project sometimes this communication can be misinterpreted. Due to high volume of interactions between different groups of stakeholders development process may misinterpret customer needs, and finally end up with a product that does not fulfill contractors' expectations. This is why DSL platform uses a specific language dedicated to describing business problem domains. Having a model discussed and represented using the unified language with unified meanings and understanding of concepts, team communication is significantly improved, resulting in a software that meets user need better. Documentation that is generated in this manner better specifies the software project, promotes consensus among team members and has overall higher quality. DSL Platform takes the documentation even one step further, since the

documentation itself represents a full formal system specification that can be readily used for rapid prototype system validation.

B. Improving efficiency of source code and automation

The formal specification of the business system can be used as a solid basis for improvement of code generation and manipulation. Dedicated compiler of DSL Platform can use this formal description of functional specifications to create any of the components for the finalized business software system. These can be libraries targeted for a particular programming language or framework or database artifacts for any relational or object-oriented database system. During the creation of the software artifacts, due to formal specifications, additional improvements of code can be automatized creating faster and more reliant execution of system tasks as well as creating more maintainable source code for the project. Finally a number of database maintenance and administration tasks can be performed using DDD model and then implementing them by simply migrating changes into a particular database system.

III. ESTIMATION OF SOFTWARE VALUE

In this Section we discuss the requirement and motivation for precise estimation of software value and describe one of novel concepts to strategically determining the value of software assets.

In strategic management one of the most important basis for decision making is the assessment of economic value assets. Even more importance for appropriate decision making is the precision in assessing the economic value of intangible assets as their value may be harder to realistically judge.

In software industry this is the case with software assets. Majority of assets are internally developed software systems that are used either to offer services on the customer markets or to sell the software itself on the customer market.

Software as an asset has some of the properties that differentiate it from any other asset, tangible or not [11]:

- 1) Indestructibility. Using software over time does not degrade its quality notwithstanding the length of usage or number of uses. Consequently this property reinforces the internal quality of software asset and its durability, so that the change in its value is solely determined by external factors. In this respect software value may deteriorate over time [13], especially with the technological advancements that change the working environment of the software.
- 2) Transmutability. Personalization, customization, modification and other altering practices of existing software systems are easily achieved which results in cost-effective production of software variants. This is particularly important for customer segmentation and price discrimination market targeting strategies [12].
- 3) Reproducibility. Since high-quality copies of the original software can be produced at low cost may authors agree that the marginal cost of production is almost zero [14]. Structure of production cost for software products contains

primarily fixed cost for the software provider. Production of each additional unit does not significantly increase the total cost. In this respect the potential reproducibility deliver to software assets also significantly improves its value.

Along with this features software assets may take advantage of different economics phenomena that can also influence the estimation of its value. We will mention just a few examples. The network effect that the use of final product or services may produce in the targeted market segment can create lock-in effects promoting customer loyalty and stabile customer base. The wider the customer base the more valuable software asset becomes according to Metcalf's law. Consequently the value of customer product and services that are based on that software asset increases proportionally. Distribution of software using corresponsive Internet services reduces or even eradicates the costs of logistic and inventory. Internet services also may transform software products into services. Many desktop applications now are available as online services (SaaS) that allow for more effective pricing strategies through pricing discrimination.

All of the above features of software assets should be taken into account during the estimation of software value.

Currently, software value estimation in practice is based on three possible approaches [15]: (1) cost-based; (2) demand-driven or value-based and (3) competition-oriented.

The cost-based approach is widely used as it is covered by the International Accounting Standard 38 – Intangible Assets (IAS 38). Main purpose of IAS is to standardize financial reports for all countries that accept the standard in order to make their financial statements comparable, basic accounting principles are adopted. For asset measurement this means that there is a preference for underestimating the asset value rather than overestimate it. This is why most of the value estimates are based on historical value which is usually lower than current value, or market value, especially for intangible assets.

Computer software is treated as an Intangible asset as it is a non-monetary asset, without physical substance and identifiable. Standard defines that its value is initially measured with cost, subsequently measured at cost or using revaluation model. Also, it takes into account future economic benefits that the asset may yield. Even though these benefits may significantly influence the value of software assets, they are usually overlooked in practice, so that during the estimation of software asset only production costs is taken into account. Even production cost does not necessarily translate into software value, since during the development of software a number of software functionalities may be developed that never make it into the final product [2], or increase in project costs that do not directly increase the value of software being developed (i.e. expensive overheads, accommodation and travel costs for team members, etc.). Poor project management practices are not taken into account during current estimation approaches as well as the quality level of software asset. All these elements may lead to overestimation of software assets

which in turn is contrary to basic accounting principles.

Accounting value used for financial reporting, therefore, does not reflect the true potential of software assets, honoring the specific properties that we described earlier, for the purpose of strategic decision making. Using accounting value will either underestimate or overestimate capitalization on the balance sheet or inevitably misrepresent due diligence before possible acquisitions. Strategic decision making requires better estimation of the potential of software assets that takes into account specific properties and potential software assets offer.

This is why new approaches are developed in order to make the estimation of software value more reliable. In the remainder of this Section we will present an estimation model based on the notion of technical debt and interest as described by Groot et al.

A. Software Valuation based on Technical Debt and Technical Interest

Technical debt is a type of opportunity cost defined as a set of quality issues or problems in software that will cost the organization that owns the software greater expenses if they are not resolved [16]. Furthermore, there are two major components of technical debt [18]:

- 1) principle, as cost to repair a software system in order to achieve ideal level of quality and
- 2) interest, as additional maintenance cost due to the lack of quality.

Technical debt increases over time if the quality issues of software are not resolved due to maintenance costs that increase as additional effort to negotiate quality issues is called for [17]. According to financial economics principle of technical debt is a cost that increases over time by the rate of interest (Figure 3).

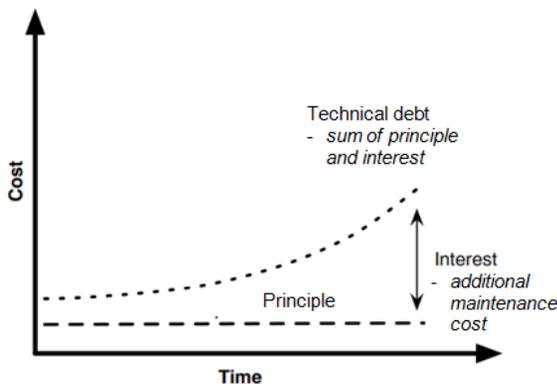


Fig. 3 Structure of Technical debt over time

Due to this increase of technical debt over time, it is feasible to pay the initial cost to repair software system and bring it to the ideal level of quality. At this level lower maintenance cost are required for the operation of the system in the future. In Figure 4 we can see that future benefits from software system operating at the ideal level of quality yielding significant savings.

In order to include technical debt in the estimation of

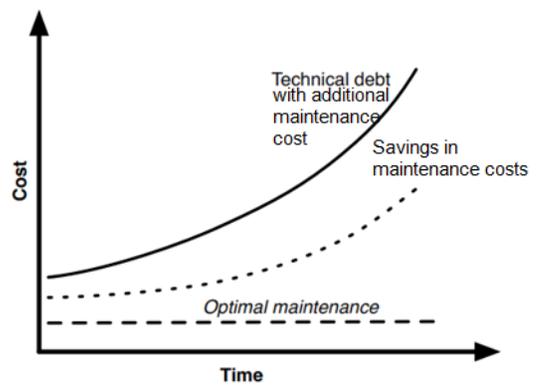


Fig. 4 Benefits from maintaining software system at the ideal level of quality

software value [2] have proposed a layered Software Valuation Pyramid model. This model relies on SIG Maintainability model (SIG) to determine the software development level and conclude the ideal level of software quality. On top of development level estimates they propose metrics that help estimate the operational costs of developed software systems with three key measures: rebuild effort, repair effort and maintenance effort (Figure 5).

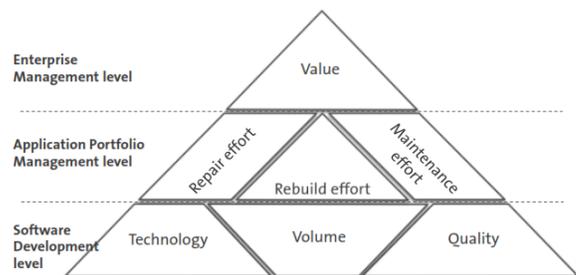


Fig. 5 Software Valuation Pyramid (Groot et al, 2012)

Rebuild effort (RbE) is defined as technology-neutral measure of technical volume, based on the technology used and volume of produced source lines of code (SLOC). Repair effort (RpE) is equal to the technical debt of the software system which is primarily determined by the quality of software development process. This means that only a part of the software system needs to be rebuilt and this part is referred to as the rework fraction (RF). Maintenance effort (ME) is the yearly effort estimated to be required for regular maintenance of the system, including bug fixes and small enhancements.

Based on the above defined metrics [2] propose tree different models of estimating software asset value.

B. Software Asset Estimation Models

For the purpose of this paper we will consider three models of estimating production value of software assets, which will be bases of analyzing impact of DDD approach to software asset development. All of the models are based on the assumptions that (1) there is a known level of software asset quality based on SIC metrics described earlier and (2) there is

an ideal level of quality for software asset at hand that is higher than the current level of quality as previous empirical studies suggested. Even if the ideal level of quality is lower than the current level of quality these models of value estimations may apply.

First model is based on Repair effort (RbE). According to this model estimated value V is equal to rebuild effort discounted by the repair effort (RpE) required to bring the quality of software asset to ideal level.

Second model is based on the Rework fraction (RF). If bringing software system to ideal level requires the replacement of complete component or set of components that the estimated value of the system V is equal to the value of the part of the system that does not require any improvements (i.e. the value of the fraction that ought not to be reworked).

Third model is based on Technical interest. Here rebuild value (RV) is discounted by the value of technical interest during the working lifespan of the software system. Technical interest is the increase of maintenance cost that occurs if the system is running in its current level of quality. The amount of additional maintenance cost is given in Figure 4 as dotted line, representing the possible increase of present value of software system if it were upgraded to its ideal level of quality before its introduction into production phase.

For further details refer to the paper [2].

IV. ANALYSIS OF SOFTWARE MANAGEMENT PRACTICES AND DOMAIN DRIVEN DESIGN

A. *Relating Software Asset Value Estimation and Software Development Approach*

As we can see in the proposed models of estimating value of software assets, all of them heavily rely on the costs that the exploitation of software asset incurs. Therefore, we may infer that software assets that are not used tend to lose their value, since there are no maintenance costs except storage costs. The value of these assets decreases until it reaches the value of acquisition as defined in IAS 38.

For software assets that are activated and operational in the production system, estimation of its value can be executed using described models. The main determinant of the estimation level will be related to the quality of software development approach. This is inevitable as the Rebuild effort (RbE) relies not only on the volume of the system (i.e. SLOC) but also the characteristics of the technology used. The technological measure includes the properties of software development environments, programming languages and practices, as well as project management principles and software approaches which results in corresponding level of software quality.

On the other hand Repair effort (RpE) takes into account the maintenance costs that heavily rely on the chosen software approach to software development life cycle (SDLC).

All of the three models benefit from the efficient software approach as the estimated value of software asset increases. If software approach allows for higher technological coefficient

the final RbV will be higher resulting in higher value estimates.

In the first model lowering the Repair effort estimate also increases the value of the value estimate. Since RpE is equal to technical debt we can see that more efficient software approach such as DDD results in increased value estimates of software asset.

In the second model lowering the Rework fraction RF increased the value estimate. This means that if more optimized source code is used smaller part of it will have to be reworked in order to increase its quality.

Finally, in the third model it is even suggested that if more efficient software development approach is adopted in later stages of software development life cycle (SDLC) it may partially improve software value of the system, as the technical interest will be discounting the rebuild value RV at a lower rate.

All of the described models can be applied to complex software systems that are composed of various development frameworks, programming paradigms and languages, database frameworks and technologies. Interconnecting this type of complex systems generates substantial additional development and maintenance costs.

If these connections can be negotiated from a single centralized programming concept represented by a unified model of the complete system the effort required to maintain the system would decrease. This is why the approach to complex software system using domain driven design may effectively influence the value of complex systems and software assets. This influence can be observed during the early development stages, but also during later stages i.e. during the production stage and maintenance of the system.

As we described earlier, DDD is focused on describing the domain. For complex systems (such as business software systems) this means that only business processes have to be described without the concern with technical details.

Business experts can communicate their understanding of business processes to system development teams using a unified ubiquitous language that also represents the formal specifications of the system. In the end, model represents the business domain at hand, with no regard to what part of the complex system it refers to (particular functionalities, external systems and data sources or databases).

Further tools that draw on DDD approach can use this formal descriptions and using compilers dedicated to particular properties of the model create system components in a flexible and yet automated way, producing optimized and maintainable source code resulting with increased software quality.

Particularly, tool DSL Platform contains a number of compilers that translate the source code of the DDD model into different segments of coupled complex heterogeneous software systems, building on top of various frameworks, languages, libraries and platforms. In this way it synchronizes the complete systems and migrates data between database and the model and vice versa. Workload for the development team

is alleviated so that team members can spend more time on designing the domain model itself in cooperation with business experts.

The disadvantage of introducing DDD in software development is the additional effort required to adopt this software development approach. As software system grows alternative software development approaches usually tend to increase maintenance cost and decrease quality of code and the system gradually degrades. With software system growth DDD establishes better management over the complexity of system with little degradation of system quality making initial entry cost feasible. Also, additional effort and time is needed to create a substantial model of the business domain before positive effects on the development process become apparent.

Benefits from moving the focus of the development team from technical issues to business logic, as well as the improvement of the communication between team members improves the quality of software systems developed. Additional saving obtained through lower maintenance cost and increased quality of source code through better performance of execution and improved manageability of code can significantly improve the value of complex business software systems. However, DDD does not seem to be widely spread and accepted in practice.

B. Investigating DDD Adoption Limitations in Software Management Practices

In order to verify the findings in this paper, several interviews were conducted with various team members from two software development companies and two financial institutions that develop their own software solutions. Based on the responses gathered during interviews SWOT analysis was conducted. Results are given in Figure 6.

SWOT matrix	advantages STRENGTHS	disadvantages WEAKNESSES
Internal	<ul style="list-style-type: none"> • better team communication • focus on business logic • automation of particular development & maintenance tasks • unified domain model • increased level of quality • increased software value 	<ul style="list-style-type: none"> • high entry costs • cost inefficiency for simple software systems • top management resistance • high level of isolation and encapsulation in domain model may present a challenge for business domain experts
External	<p>OPPORTUNITIES</p> <ul style="list-style-type: none"> • improved estimation of value for developed software assets • reduction of maintenance costs during production phase of software system • prolonged lifespan of software systems • sustaining business logic of legacy systems 	<p>THREATS</p> <ul style="list-style-type: none"> • incentive to maintain legacy technologies and programming languages while maintaining high software value • as changes in domain model are reflected in system components risk of human error increases

Fig. 6 SWOT analysis of DDD approach to complex business software systems

The advantages were concluded based on the evidence described in this paper while the disadvantages needed further

assessment and data collection obtained through interviews. Interviews were largely used to identify weaknesses and threats of adoption DDD approach for development and maintenance of complex business systems.

As we can see in Figure 6 strengths refer to core advantages of DDD with high emphasis on software management issues and especially business management aspects of software management, such as focus on business logic, unifying business domain for all team members regardless of their background and benefits in software quality and, particularly important for in-house development, increased software asset value.

On the other hand weaknesses of adopting DDD pertain to initial cost of adopting this approach as well as the risk of overestimating final system complexity as DDD is highly cost inefficient for simple software system.

The most important weakness is the current state top management awareness which represent the main limitation to wider adoption of this approach. The highest benefits can be achieved in large-scale non-software companies that develop in-house software solutions, such as financial institutions and banks, where the focus of core business is not on software development. These are also the companies where awareness and understanding of potential benefits seems to be at a comparatively low level as well as the priority in managing software development approaches. The main obstacle preventing the higher acceptance of the domain driven design in practice is the lack of understanding the benefits of DDD and potential tools it provides by top level management. As the bottom-line in risk management is to prevent potential risks, additional adjustments of value estimations of software systems does not justify adoption of DDD in companies that were interviewed. Additionally, successful adoption requires business domain experts to adjust to the domain specific language which is characterized by high level of isolation and encapsulation which is more familiar to software experts.

External elements of the SWOT analysis describe the potentials of adopting DDD where positive potentials represent opportunities to be gained. As we can see in Figure 6 improved valuations of software assets can be achieved and in turn promote better strategic decision making. Also, reduction of maintenance cost during production phase improves internal rate of return on investment while at the same time extending the lifespan of software asset. Equally important is the potential of preserving business logic in legacy systems which would be otherwise either lost after the discontinuation of legacy systems or retained through expensive process of reengineering.

Prolonged lifespan may also lead to one of two most important threats in adopting DDD. This is the incentive to maintain legacy systems that rely on old technologies, programming languages, paradigms or frameworks while maintaining high software asset value which may expose the company to additional risks such as self-exclusion from trends in software developments and increase of inefficiency resulting

in loss of competitive advantages. Additional threat that can be detected is the possible increase of the importance of human error factors since the software model is directly related to the system itself, so that any change is readily implemented in software components in the production phase.

V. CONCLUSION

In this paper we have presented one of more recent approaches to software development called domain driven design (DDD). We assessed its implications on software management process through impact on software value estimation and changes in maintenance efficiency. As this approach is still to see its wider adoption in practice we first took a look at its main characteristics and, building on current research, position it according to recent classifications. By comparison with other approaches we classified DDD to an intermediary group between structured approaches and behavioral approaches. In fact, DDD seems to have been the missing link since the intermediary class only recognized a class of methodologies based on agile software development concerned with small and medium projects. DDD completes the classification as it is intended for complex heterogeneous software systems.

For the purpose of this paper we took two main benefits from DDD describing their practical implementations through an existing tool DSL Platform. We estimated the impact of these features on two major issues in software management – software value estimation and maintenance cost effectiveness. We have shown that level of quality of software can be greatly improved during development phase through better communication and moving focus from technical to business arena. During the production phase of software system higher quality of code optimizes maintenance cost in comparison to suboptimal software system quality. All of this is reflected through software asset value. We have shown building on software valuation models presented by Groot et al (2012) how the changes DDD provides impact all of the three proposed models of software valuation.

Finally we have conducted interviews with information officers and managers in software companies and banks to obtain data and create a SWOT analysis of adopting DDD in companies that manage in-house complex heterogeneous software assets. The analysis showed that main obstacle for adoption of DDD is lack of understanding the economic benefits by the top management.

This is an important confirmation of current limitation to adoption of DDD in mainstream software industry and software departments of large companies that should be taken into account when communicating research information to business users and management.

ACKNOWLEDGMENT

I would like to thank Rikard Pavelic and company Nova Generacija Softvera d.o.o. for their cooperation during research of this topic, donating free access to DSL Platform

(<http://dsl-platform.com>) for the purpose of evaluation and invaluable information that improved the quality of the research results presented in this paper.

REFERENCES

- [1] N. Mavetra and J. Kroeze, “Guiding Principles for Developing Adaptive Software Products” in *Communications of IBIMA*, vol. 2010, IBIMA Publishing, 2010, pp. 1 – 15.
- [2] J. de Groot, A. Nugroho, T. Back and J. Visser, “What is the value of your software?” in *Proceedings of the Third International Workshop on Managing Technical Debt (MTD)*, 5th June 2012, Zurich: IEEE, 2012, pp. 37–44.
- [3] N. R. Jennings, “On Agent-based Software Engineering” in *Artificial Intelligence*, vol. 117, Elsevier Science, B.V., 2000, pp. 277 – 296.
- [4] D. Sharma, W. Ma, D. Tran and M. Anderson, “A Novel Approach to Programming: Agent Based Software Engineering” in *Knowledge-based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science*, vol. 4253, Berlin: Springer Verlag, 2006, pp. 1184 – 1191.
- [5] D. North, “Behavior Modification: The evolution of behavior-driven development”, in *Better Software*, vol.-issue 2006-03, Techwell Corp.
- [6] R. Brown, S. Nerur and C. Slinkman, “The philosophical Shifts in Software Development” in *Proceedings in the 10th Americas Conference on Information Systems*, New York, August 2004, pp. 4136 – 4143.
- [7] E. Evans, *Domain-Driven Design: Tackling Complexity in the Heart of Software*, Addison-Wesley, 2004.
- [8] G. Booch, J. Rumbaugh and I. Jacobson, *The Unified Modelling Language User Guide*, 2nd Ed., Addison-Wesley, 2005.
- [9] J. S. Cuadrado and J. G. Molina, “Building Domain-Specific Languages for Model-Driven Development” in *IEEE Software*, vol. 24, Issue No. 5. IEEE Computer Society, September/October 2007, pp. 48 – 55.
- [10] R. J. Wirfs-Brock, “Driven to... Discovering Your Design Values” in *IEEE Software*, vol. 24, Issue No. 1. IEEE Computer Society, January/February 2007, pp. 9 – 11.
- [11] S. Y. Choi, D. O. Stahl and A. B. Whinston, *The economics of electronic commerce: the essential of doing business in the electronic marketplace*. Indianapolis: Macmillan, 1997.
- [12] S. Lehmann and P. Buxmann, “Pricing Strategies of Software Vendors” in *Business & Information Systems Engineering*, vol. 6, Heidelberg: Springer Verlag, 2009, pp. 452 – 462.
- [13] J. Zhang and A. Seidmann, “The optimal software licencing policy under quality uncertainty”, in *The Proceedings of the 5th international conference on electronic commerce*, New York: ACM Press, 2003, pp. 276–286.
- [14] S. Royer, *Strategic Management and Online Selling: Creating competitive advantage with intangible web goods*, New York: Routledge, 2005.
- [15] C. Homburg and H. Krohmer, *Marketing Management: Strategy – Instruments – Implementation – Governance*, 2nd Ed. (in German), Wiesbaden: Gebler, 2006.
- [16] W. Cunningham, “The WyCash portfolio management system,” *ACM SIGPLAN OOPS Messenger*, vol. 4, no. 2, 1993., pp. 29–30.
- [17] A. Nugroho, J. Visser, and T. Kuipers, “An empirical model of technical debt and interest,” in *Proceeding of the 2nd International Workshop on Managing Technical Debt.*, ACM, 2011, pp. 1–8.
- [18] B. Curtis, J. Sappidi and A. Szykarski, “Estimating the Size, Cost, and Types of Technical Debt”, in *The Proceedings of the International Workshop on Managing Technical Debt*, 2012, Zurich, Switzerland.

Comparative Advantages of Software Industry in Developing Countries: Study of Structure, Market Strategies and Software Development Approaches in Croatian Software Companies

Nikola Vlahovic, Ljubica Milanovic Glavan and Anja Frankovic

Abstract—During the global economic crises national software industries have proven to be one of the most resilient industries. After initial fall of the market in 2008 and 2009, recovery followed with a period of intensive innovation. Some Asian software industries even reported no influence on the market share, such as Thai software industry. Nevertheless in smaller economics impact was more pronounced but in comparison with other national industries the recovery is generally faster.

In this paper we will concentrate on software companies in developing countries and try to investigate the main characteristics of small and developing software industry that may create resilience in difficult economic setting and serve as a basis for international competitiveness. For the purpose of this paper research was conducted on the case of Croatia.

The goal of this paper is to detect the most important opportunities and sources of comparative advantages that open these companies in global international software markets. Based on the conducted research of software industry structure, market targeting and software development practices guidelines are outlined on how to strengthen the competitive advantage of small national software industries in developing countries on the global software markets.

Keywords—Agile software development, Cloud computing, Comparative advantages, ICT industry, Offshoring, Software development methodologies, Software industry.

I. INTRODUCTION

RECENT global economic crises have witnessed that some industries suffer significantly more from fluctuations in global financial and economic markets during recession than others. One of the most resilient industries has proven to be industries belonging to the ICT sector. Some of these

N. Vlahovic is the associate professor at the Informatics Department of the Faculty of Economics and Business, University of Zagreb in Croatia. Trg. J.F. Kennedyja 6, 10000 Zagreb, Croatia (phone: +385-1-238 3220; fax: +385-1-233 5633; e-mail: nvlahovic@efzg.hr).

Lj. Milanovic Glavan is postdoctoral researcher and assistant at the Informatics Department of the Faculty of Economics and Business, University of Zagreb in Croatia. Trg. J.F. Kennedyja 6, 10000 Zagreb, Croatia (e-mail: ljmilanovic@efzg.hr).

A. Frankovic is bachelor of Economics who graduated the Managerial Informatics Study Program at the Faculty of Economics and Business, University of Zagreb in Croatia. Trg. J.F. Kennedyja 6, 10000 Zagreb, Croatia (e-mail: anja.frankovic@gmail.com).

industries have recorded smaller rates of decline in revenues than other industries, while others reported no influence whatsoever. A good example is Germany where after the initial decline during 2008 and 2009 software and IT industry recovered in short period of time. Thai software industry reported no influence of the crisis as it readjusted its strategies and turned successfully to Asian markets for support [19]. European commission has even relied on these industries (particular consumer electronics, gaming industry, and telecommunications and high technology sector) to lead the path to economic recovery believing that these industries can induce higher personal consumption expenditures.

Developing countries also experienced similar effects on their national software industries. As the national economic stability is primary goal especially in developing and transitional countries of South Eastern Europe investigating the foundations of this comparative advantage in comparison to other industry sector was prompted.

In this paper we will try to investigate specific features of software industry pertaining to structure, market strategies and software development practices and determine how these features contribute to this comparative advantage.

Goal of this paper is to provide the insight in current software development practices that allow small and medium companies, which form the major share of software industry in developing countries, to compete successfully on national and international software markets providing them with additional resilience against local and global market instabilities. Main focus will be on the research of diversification in market coverage and internationalization, as well as innovation since current research indicates that these elements are most important characteristics of software industries that resisted oscillations during global economic crisis.

Structure of this paper is as follows: in Section II we will present background on specific characteristics of global software industry that distinguishes it from other industry sectors. Here special attention will be given to the product and services this sector provides as well as the overview of latest methodologies used in software development. In Section III we will make an overview of recent developments across global software industry and various software industries in

Europe. We will define international economic setting, software markets and software trends. Here the scope of research pertaining to small national industries of developing transitional countries will be defined. In Section IV we will describe conducted research in Croatia, defining the methodology used and obtained results describing the main features of software industry in Croatia. In Section V discussion of results in relation to current trends will be given. We will make a comparative analysis of the research findings in Croatia with software industries in other developed and developing countries in Europe. Finally in Section VI we will present the conclusions and indicate future trends and possibilities of software industry in developing countries.

II. BACKGROUND ON SOFTWARE INDUSTRY FEATURES

Software industry along with information and communication technology is one of the most important sectors of international economics in the information age. Product and services that are provided through this economic sector are of essential importance to overall economic development, business development and scientific research. It provides an essential incentive to both business spending and personal consumption expenditure which in turn generate economic growth. On the other hand software industry has stayed elusive to industry analyzers and financial markets longer than any other sector [13], [14]. This is primarily due to a high level of variation that this industry provides to the constraints exhibited by analytical methodologies and their categorizations. Furthermore this industry is characterized with one of the highest levels of innovation resulting in creation of distinctly varied products and services spanning over not only software industry but also spilling over and expanding to numerous other industries in various forms and approaches. Yet software industry has anchored itself as a unified industry proving to be highly stable and resistant to economic fluctuations and market instabilities.

Some of the most important characteristics of the industry include the fact that its main products are the most complicated man made products and yet intangible in its nature. This fact alone calls for highly skilled professionals that open up a demand for highly trained professionals such as computer programmers, software architects and software designers. In turn, high expectations and use of hi-technology in their work is compensated by above average salaries that for over a decade have positioned these job positions in highest paid job positions rankings. Additionally, with the development of Internet and Internet Services, particularly the World Wide Web service, and due to intangible and digital nature of software products, additional access both to job positions and customer markets was alleviated and raised to international and global level. Currently the most important characteristics that drive the growth of software industry market share along with the innovation potential it offers heavily rely on the definition and repositioning of software products from products to services and efficient organization

of software development process.

A. *Software as a Product or Service*

Software is different from all other types of goods due to its intangible nature. By definition software represents the unity of computer programs and their documentation [1] which allow the computer system to operate and be used by their end users. Computer programs represent an organized set of digitalized instructions intended for computer systems that allows them to perform specific tasks for their users [2]. Software documentation describes program functionalities. It consists of technical documentation and user documentation that is usually part of the software solution presented through user interface during software execution but also in separated form as external resource. Software can also refer to other intangible assets of computer systems, such as database definitions and models and data contained in these systems, as well as various protocols used in computer communication.

Software is a digital good and as such, from an economic perspective it has three fundamental properties: indestructibility, transmutability and reproducibility [3].

- 1) **Indestructibility.** Using software over time does not degrade its quality notwithstanding the length of usage or number of uses. This may lead to conclude that value of created software does not change, but external influences have a decisive influence on software value. In this respect software value may deteriorate over time [4], as technological advancements change working environment of existing software solutions.
- 2) **Transmutability.** Personalization, customization, modification and other altering practices of existing software systems are easily achieved which results in cost-effective production of software variants. This is particularly important for customer segmentation and price discrimination market targeting strategies [5].
- 3) **Reproducibility.** Since high-quality copies of the original software can be produced at low cost may authors agree that the marginal cost of production is almost zero [6]. Structure of production cost for software products contains primarily fixed cost for the software provider. Production of each additional unit does not significantly increase the total cost. In this respect the potential reproducibility deliver to software assets also significantly improves its value.

Chronologically, through these fundamental features of digital goods, software evolved from a particular product delivered recorded on a material medium (such as data disks) towards less materialized forms. This transition was driven by the potential of creating additional value for customers and comparative advantage for its developers. Through habituation of end users developers were able to take advantage of lock-in effects in targeted market segment and build a solid foundation of stable customer base.

In order to achieve these potential advantages software companies started taking advantage of digital distribution of

software using responsive Internet services reduced or even eradicated the costs of logistic and inventory, while additionally promoting convenience for customers, allowing for automatic update of software components.

Finally, dematerialization of software distribution and the further development of Internet's communication capacities lead to creation of cloud computing where software itself is not even delivered to users as a compact set of computer programs and components. Software is radically being transferred to the cloud paradigm and being offered as a service. This transition was also gradual as we can see in Figure 1.

Cloud Service Models

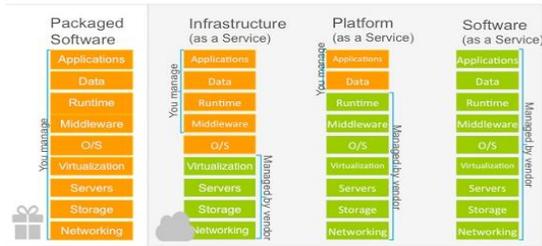


Fig. 1 Comparing stand-alone applications and Cloud computing services

Providing Infrastructure as a Service (IaaS) opened up new opportunities for network and internet provider businesses, providing Platform as a Service (PaaS) opened up new opportunities for software developers and providing Software as Service (SaaS) offered additional benefits for end users.

This radical change has raised the dynamics in the software markets, bringing new concepts for both Internet providers and software developers in terms of decreasing the investment and maintenance costs while improving efficiency. In turn end customers benefit from lower prices for their information needs, more fair pricing models and higher robustness and longevity of their data as ubiquitous computing concepts become implemented.

All of these trends have been recognized not only by large multinational companies but also smaller entrepreneurs that recognized the opportunity to acquire powerful tools and start innovative businesses.

B. Current Software Development Methodologies

With the increase of dynamics in software markets and improved availability of tools changes in approach to developing software were immanent.

Additional pressure in creating fast quality software solutions promoted the use of agile software development methodologies.

There are twelve principles of agile development, as defined in Agile Manifesto [7] that describe the values and standpoint in which agile methodologies for software development should be founded. Primarily, agile approaches focus on individuals instead on the process while promoting improved communication among team members and other stakeholders. Key is the development of working software that can be easily and quickly changed to adapt to changing user

requirements and dynamic environment. Some of the most important agile methodologies include Scrum, eXtreme programming (XP), Kanban, Feature-driven development, etc. [8]. Limitations that all of these methods share is the size of software system developed since agile approach works best with small development teams and systems with limited complexity. Large software systems developed by large development teams which may rely on legacy infrastructure cannot directly benefit from agile approach. This is way combinations of more traditional approaches such as sequential 'waterfall' development, incremental development approaches or unified process approaches, with agile methods tried to implement agile principles to larger complex software systems. Resulting methodologies have been implemented, such as Agile Unified Process (AUP), Scaled Agile Framework and Large-scale Scrum.

Along with Agile development different practices pertaining to one or more of its methodologies have encouraged the development of more recent tools that are based on approaches such as Behavior-driven design (BDD), Domain-driven design (DDD), Continuous Integration (CI), etc.

Still, for legacy systems and large-scale complex heterogeneous system traditional structured approaches are still widely used.

III. RECENT SOFTWARE INDUSTRY TRENDS IN EUROPE

European software market is second largest software market in the world with more than 231 billion EUR of global market share [9]. It is ranked after Northern America and ahead of Asia & Pacific Region (including Japan) (Figure 2).

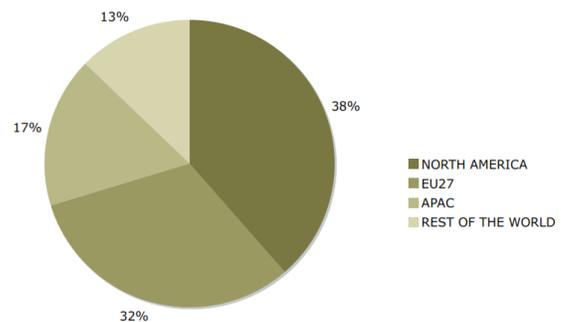


Fig. 2 Global Software Markets by Region

This is the main driving force for software industry that has shown positive trends over the last decade, despite global economic fluctuations and occasional economic crises. Appropriately, R&D spending in software industry is also second largest in the world (after United States and Canada), increasing year after year. In the last five years, though, the rate of market grow is decreasing and profits have fallen below the R&D spending (Figure 3).

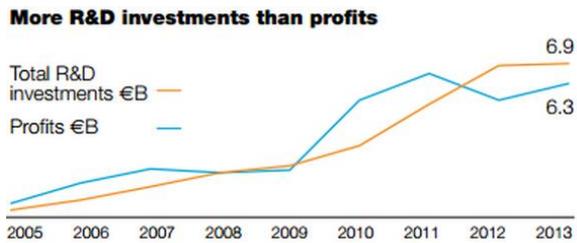


Fig. 3 R&D spending and profits of EU software industry (Source: Truffle Capital)

The main contributor to the EU software market is Germany that accounts for half of the revenues, followed by United Kingdom with 13% revenue share, France 12% revenue share, Sweden with 6% revenue share, The Netherlands with 5% revenue share, and other countries with less than 3% share each [10].

For the most part, software industry along with accompanying ICT sectors was the driving force of resisting the restrictive fluctuations and economic recession during 2008 and 2009. During that period software market lost about 5% but recovered quickly in the following years [11]. This is the exact timeframe for increased innovative spur that took the advantage of new trends in ICT and software. Some of the most important innovative products and services include business intelligence software, IT security software, Enterprise Content Management, SOA and software as a Service (SaaS). Dominant number of innovations relied on cloud computing.

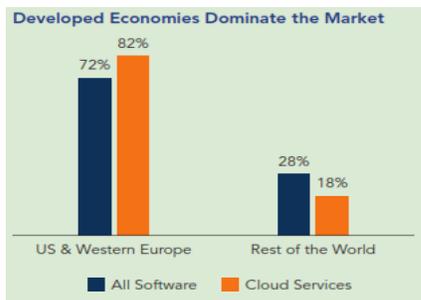


Fig. 4 Global software market shares between developed and developing countries

As we can see in figure 4 globally, developed countries intensified development of cloud based services while developing countries followed. The same trend is seen in European region.

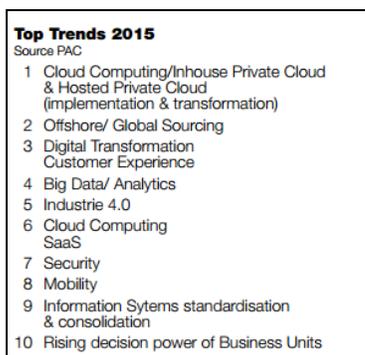


Fig. 5 Top trends in software industry in 2015

As cloud computing constituted only 9% of global software market in 2013 [12] currently it still dictates leading trends in terms of IaaS/PaaS and SaaS challenging the potential for innovation in software industry (Figure 5).

In the past decade developing countries have taken the opportunities for offshoring of software by the leading developed countries [15], [17]. For East and Central European countries this trend was additionally enhanced as they had the opportunity to develop the required information infrastructure using financial support of the EU funds.

Benefits are mutual for both developed and developing countries. Cost of software development reduces for the leading software industries of developed countries while creation of new software producing regions promotes overall economic development in developing countries. It has been also shown that developing countries that were initiating growth of their software industries through direct foreign investment, such as Estonia and Romania, have been able to establish and develop software industries more successfully than countries that relied primarily on domestic investment, such as Bulgaria [16].

These investment circumstances and open possibility for offshoring developed countries in software development tasks provided additional resilience to national economic fluctuations observed in the past 5 to 8 years in transitional countries.

Still features of developing countries' software companies that allow for offshoring and comparative advantage remain unclear, insufficiently investigated and defined. In the rest of the paper we will present a study on the case of Croatia trying to establish the main features of software industry in developing transitional country.

IV. RESEARCHING COMPARATIVE ADVANTAGE OF SOFTWARE INDUSTRY IN DEVELOPING COUNTRIES: A CASE OF CROATIA

In order to understand better the competitive position of software industries in developing countries a research was conducted in Croatia. Subject of the study was focused on active companies that are according to Croatian chamber of economy registered for activities that include computer programming, consultancy and related activities and information service activities services. Also, the study included only companies that have reported a minimum of 12.000 euros of revenue in 2013.

Goal of the survey was to determine the demographics of the companies included in the research, understand the market they are selling their products and services and finally understand their internal organization with additional emphasis on software development methodology that is prevalent in their business operations.

A. Research Methodology

The research was based on acquiring data from secondary sources (i.e. Company Register of the Croatian Chamber of Commerce), and primary source through a survey dedicated to the earlier described goals of the study.

Survey was organized in three sections. First section contained question pertaining to the demographics of companies including size of company, revenue, geographical location and number of employees. Second section of the survey was used to investigate target markets and the ratio between domestic consumer markets and international market segments. Finally, in the third section of the survey information about the methodology used in during the software development life cycle was collected.

The survey was sent electronically to 650 companies, addressed to middle-level management. Exactly 100 of companies sent their responses which equals to 15,4% response rate.

Responses were then processes and descriptive statistical analysis was conducted.

B. Research Results

The average age of companies is 12,6 years as the registration of companies first began in 1990 when Croatia gained independence. Some of the companies registered in first couple of years were actually active even before, but overall we can say that software industry in Croatia is a young industry with average rate of growth of 4 companies with yearly revenues of at least 12.000 EUR each year.

Software companies gravitate towards Zagreb city area as this is the major financial and business area in the country. 63 companies are registered in Zagreb while the others are registered in other cities.

While only two of largest companies had more than 100 employees, majority of 71 companies in 2013 had less than 10 employees, while further 23 companies had between 10 and 50 employees (Figure 6).

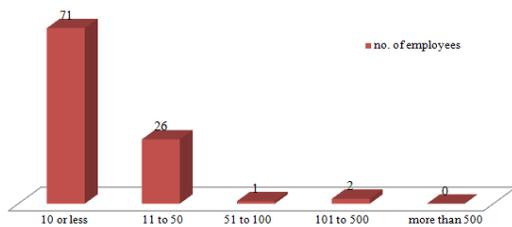


Fig. 6 Software companies by size – number of employees

45% of companies estimated that their revenue in 2014 will remain in vicinity of 100.000 EUR or less.

In Figure 7 we can see that more than 60% of companies produce software i.e. computer programs, 18% provide services in maintenance of computer systems and equipment.

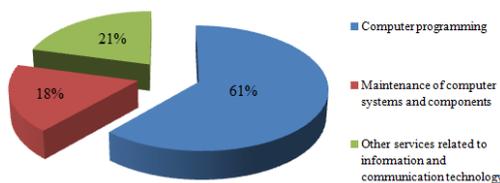


Fig. 7 Software companies by dominant business activity

The rest of the companies provide other types of information services and consultancy.

Companies earn majority of their revenue on domestic markets in 38% while remaining 62% earn their revenue abroad (Figure 8).

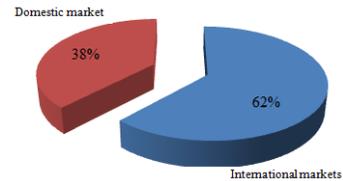


Fig. 8 Ratio of domestic and international markets

Those that export their products and services primarily export to European countries (40%) and countries of the neighboring South-Eastern countries (35%). About one fifth of export is realized on the American market (Figure 9).

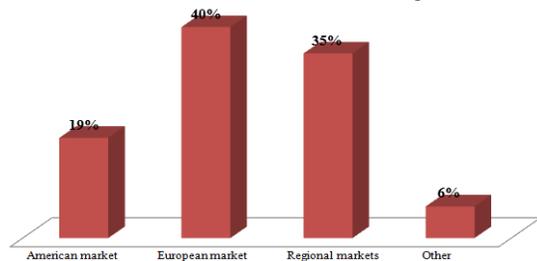


Fig. 9 Software companies exports by region

48% of companies have adopted primarily agile methodologies for the development of their software products.

While Scrum is the methodology most widely accepted, companies in most cases use a combination of agile methodologies such as combinations of Scrum, Adaptive Software Development and eXtreme Programming. There is a significant 15% of these companies that rely on the Open Source development (Figure 10).

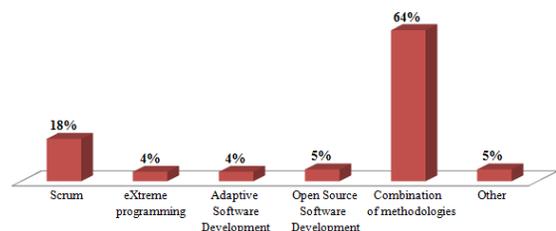


Fig. 10 Companies that adopted agile methodologies and most dominant methodologies used

Rest of the companies (52%) use more traditional approaches to software development. In this case companies mostly use component-based development taking the advantage of reusability (55%) and RUP Methodology (19%). Other methodologies that are used in lesser extent are Joint application design and other rapid software development methodologies (Figure 11).

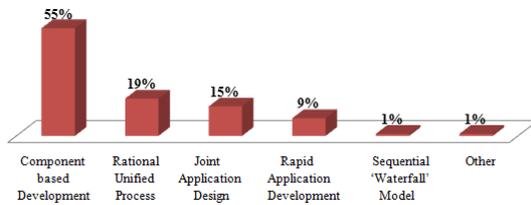


Fig. 11 Companies that use traditional approaches to software development and most dominant methodologies

There is a good representation and diversification of various architectural designs of software solutions covering desktop applications, client-server architecture and service oriented architecture (SOA), but also significant number of solutions of web applications, mobile applications and cloud computing implementations (Figure 12).

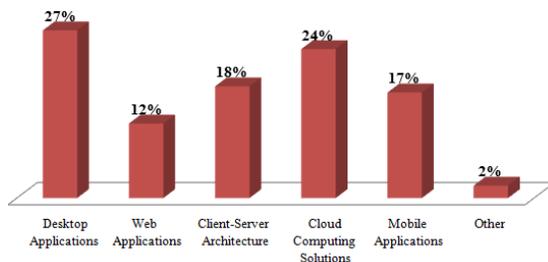


Fig. 12 Dominant Software Architecture of developed solutions

Acquired data also showed that majority, 53%, of solutions companies produce are commissioned custom made software solutions, 34% of solutions are modified Component-of-the-Shelf (COTS) and the remaining 13% belong to Open source solutions.

V. DISCUSSION

According to Croatian Bureau of statistics production of software in Croatia between 2008 and 2012 has been continually increasing following the more general trend in EU software industry, despite global economic crisis (Figure 13).

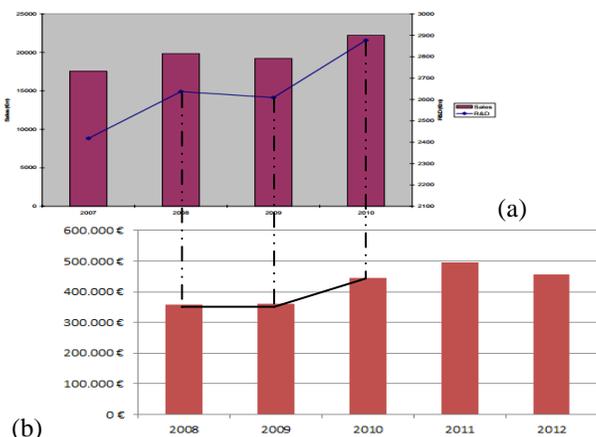


Fig. 13 Comparing trends in revenue from software in (a) EU and (b) Croatia

On the other hand during that same period Croatian GDP had negative growth rate. Hence, software industry showed remarkable resilience to negative global trends. As we can see this is also true for other developing countries in EU that we described earlier. This means that specific nature of software industry, its position and opportunities can be used to alleviate negative trends in economic fluctuations. This opportunity was not fully realized due to overall small fraction of software industry in total national GDP.

Results of the research show that Croatian software industry consists of only several large companies and a group of young small and medium companies (Figure 6). Unlike large companies, these small and medium companies have adopted agile approach to software development as their main competitive edge both in domestic but also in international markets. Fast development, quick-win strategies in discovering and targeting new markets, infrastructural readiness and familiarity with recent technologies are the main attributes that allow dynamics and readiness to follow global trends in software industry.

If we cross-reference the main findings of the research with recent trends in software industry we will see that:

- 1) One third of revenues of Croatian software companies (Figures 8 and 9) comes from two leading software markets (Northern America and Europe in Figure 2), testifying significant level of offshoring to developed economies. This strategy allows for better inflow of international capital in situations when domestic capital is scarce or too expensive. Also offshoring improves domestic knowledge, as well as it may empower domestic companies that can make use of code and components produced as by-products for initiation of propriety software projects and services. This is possible due to reproducibility and transmutability of software.
- 2) Even though packaged/desktop software is dominating Croatian production (Figure 12) as in any developing country (Figure 4), cloud computing is well represented (Figure 12) allowing for participation in global trends (Figure 5) through innovation of these services reassuring continued growth.
- 3) Agile methodologies and high number of small and medium companies make national software industry agile and prepared for dynamics typical to the global markets.
- 4) Strengthening software industry by national strategies and institutional support can help overall initiation of economic activity since all industries rely on ICT in greater or lesser degree. Taking advantage of international support through European institutions offering financing of infrastructural projects, regional development projects and scientific research projects through EU Funding may develop a long-term comparative advantage [18].

VI. CONCLUSIONS

In this paper we have described the main characteristic of the global software markets and software industries. We have indicated current trends in terms of technological innovation but also in terms of business practices and macroeconomic conditions.

Distinction between developed and developing countries and their respective software industries was made. Even though both developed and developing countries have shown a great level of resilience to economic fluctuations in performance of software industries. Developing countries are more suspect to suffer from economic fluctuations so we tried to determine key features of software industries in developing countries that make this type of industry more resilient to macroeconomic conditions than other types of industries.

Through literature overview and secondary statistical data we have shown that there is a continuous growth of software industry on any level, especially during global economic crisis in 2008 and 2009.

In this context research of structure, market strategies and software development approaches was conducted among Croatian software companies.

Results have shown that there is a significant comparative advantage of strengthening software industry in developing and transitional countries since its benefits on the entire economy are valuable. Stabilizing effect during decrease of economic activity, investment potential especially through direct international investment, and innovation are the main benefits detected during the research. The main prerequisite of obtaining these advantages is agile orientation of software companies, openness to international markets and solid infrastructure.

REFERENCES

- [1] I. Sommerville, *Software Engineering*, 10th Edition, Pearson Publishing, 2015.
- [2] R. M. Stair and G. W. Reynolds, *Fundamentals of Information Systems*, Sixth Edition. Course Technology, Cengage Learning, 2012.
- [3] S. Y. Choi, D. O. Stahl and A. B. Whinston, *The economics of electronic commerce: the essential of doing business in the electronic marketplace*. Indianapolis: Macmillan, 1997.
- [4] J. Zhang and A. Seidmann, "The optimal software licencing policy under quality uncertainty", in *The Proceedings of the 5th international conference on electronic commerce*, New York: ACM Press, 2003, pp. 276–286.
- [5] S. Lehmann and P. Buxmann, "Pricing Strategies of Software Vendors" in *Business & Information Systems Engineering*, vol. 6, Heidelberg: Springer Verlag, 2009, pp. 452 – 462.
- [6] S. Royer, *Strategic Management and Online Selling: Creating competitive advantage with intangible web goods*, New York: Routledge, 2005.
- [7] K. Beck et al., *Maniphesto for Agile Software Development*, Agile Alliance, 2001.
- [8] Versionone, *State of Agile Development Survey Results*, available at http://www.versionone.com/state_of_agile_development_survey/2011/, 2011.
- [9] Pierre Audoin Consultants, *Economic and Social Impact of Software & Software-Based Services*, 2010.
- [10] Truffle Capital, *Top 100 European Software vendors: the best software companies*, available at <http://www.truffle100.com/2014/countries.php> , 2015.
- [11] IBP, *Germany – Doing business for everyone guide: Practical information and contacts*, International Business Publications, USA, 2013.
- [12] BSA, *The Compliance Gap: BSA Global Software Survey, Autumn 2014*, BSA, 2014.
- [13] European Commission, "The European ICT industry at the crossroad: economic crisis and innovation", Pillar 5 of *Digital Agenda Scoreboard 2011* , EC, 2011.
- [14] K. Lukac, " Software project management at Republic of Croatia" in *Economic review(in Croatian)*, *Ekonomski pregled*, vol. 53, no. 1-2, Zagreb: Hrvatsko društvo ekonomista, 2002, pp.164 – 190.
- [15] A. Arora and A. Gambardella, "The Globalization of the Software Industry: Perspectives and Opportunities for Developed and Developing Countries" in *Innovation Policy and the Economy*, Volume 5, MIT Press, 2005, pp.1 – 32.
- [16] S. Mancheva, *Successful Industry Building in Transition Countries: Foreign Direct Investment or Local Effort? Software Industries of Bulgaria, Estonia, and Romania*, Saarbruecken: VDB Verlag, Germany, 2008.
- [17] D. A. Vogel and J. E. Connelly, "Best practices for dealing with offshore software development", in *Handbook of Business Strategy*, Vol. 6 Issue 1, Emerald Group Publishing, 2005, pp. 281 - 286
- [18] B.H. Rudall and C.J.H. Mann, "Emerging technologies: software-intensive systems and other current developments", in: *Kybernetes*, Vol. 38, Issue 3/4, Emerald Group Publishing, 2009, pp. 549 – 555.
- [19] A. Pomwasin, "Economic crisis no threat to software industry – ATSI" in *The Nation*, available at http://www.nationmultimedia.com/2009/01/13/technology/technology_30093086.php, 2009.

The application of computer technology in optimizing the conditions of directional breaking of fibrous collagen linkages

Shalbuev Dm.V, Zharnikova E.V., and Radnaeva V.D.

Abstract— The application of mathematical modeling in technological processes of production allows to control the process, quality and production volume, and also to predict the level of technogenic impact on the environment while creating environmentally benign technologies. Recent studies including the development of an innovative method for recycling of protein waste as well as technologies of sheepskin raw materials pickling applying collagen dissolution products (CDP) are used as a tool of mathematical modeling.

The aim of the research was the construction of a mathematical model of symbiotic effects of acids and bacteria upon the structure of the modified collagen and optimization of the environment for directed breaking of ties of fibrous collagen with the software application.

Keywords—computer technology, modified collagen, mathematical modeling, fermented milk compositions.

THE present level of science and technology development with the application of information technologies and the methodology of the system analysis allows to increasingly explore the processes in chemical engineering and microbiology with regard to the phenomena and effects of a lower level of the hierarchical structure. The analysis and solution of this problem is assumed by improving the efficiency of the use of up-to-date mathematical apparatus and methods for system studies.

The application of mathematical modeling in technological processes of production allows to control the process, quality and production volume, and also to predict the level of technogenic impact on the environment while creating

This work was supported by the East Siberia state university of technology and management under Grant «Young scientists of East Siberia state University of technology and management-2015».

Dm.V. Shalbuev is with the East Siberia state university of technology and management, Russian Federation, Republic of Buryatia, Ulan-Ude, 40V Klyuchevskaya st., 670013 (corresponding author to provide phone: +7(3012)41-72-22; +79146311809; fax: +7(3012)41-71-50; e-mail: shalbuevd@mail.ru).

E.V. Zharnikova is with the East Siberia state university of technology and management, Russian Federation, Republic of Buryatia, Ulan-Ude, 40V Klyuchevskaya st., 670013 (corresponding author to provide phone: +79246568498; e-mail: zharnikova_ev@mail.ru).

V.D. Radnaeva is with the East Siberia state university of technology and management, Russian Federation, Republic of Buryatia, Ulan-Ude, 40V Klyuchevskaya st., 670013 (corresponding author to provide phone: +79244584656; e-mail: radnaevav@mail.ru).

environmentally benign technologies.

The introduction of mathematical methods of experiment planning allows you to:

1. To exclude considerably the intuitive approach, to replace it with the scientifically substantiated program of conducting experiments, including an objective assessment of the experiment results at all subsequent stages of the research;

2. To determine the minimum number of experiments, allowing to carry out valid statistical interpretation of the results at each stage of the study;

3. Even in case of the incomplete study of the mechanism of the process it can be possible to obtain its mathematical model including the most important input parameters.

Such mathematical model can be used to control the process and determine the appropriate mode of operation.

In addition, using this model it is possible to adjust and refine theoretical ideas about the studied process [1].

The basic models of microbiology in the form of simple mathematical equations reflect the most important qualitative properties of living systems: the possibility of growth and its limitations, the ability to switch, spatial-temporal heterogeneity.

These models are applied to explore the possibility in principle of spatial-temporal dynamics of systems behaviour, their interactions, the behavior of the systems under various external influences - random, periodic, etc.

The analysis of the literature in the field of the technology of leather and fur allows to draw a conclusion about the necessity to create new methods of processing of collagen-containing raw materials, taking into account the requirements for environmental clean production.

Recent studies including the development of an innovative method for recycling of protein waste as well as technologies of sheepskin raw materials pickling applying collagen dissolution products (CDP) are used as a tool of mathematical modeling.

The aim of the research was the construction of a mathematical model of symbiotic effects of acids and bacteria upon the structure of the modified collagen and optimization of the environment for directed breaking of ties of fibrous collagen with the software application.

I. MATERIALS AND METHODS

The process of obtaining products of collagen dissolution is a complex chemical-technological processes based on the destruction of both alkali-and acid-labile ties.

To reduce the losses of protein in the process of destruction of acid-labile ties it is suggested to use sour milk composition (SMC) previously developed by the authors [2].

SMC represent a symbiosis of acid-tolerant microorganisms and organic acids and enzymes produced by them. It is assumed that the presence of acid-tolerant microorganisms in SMC will ensure the destruction of acid-tolerant ties. Based on the above it can be assumed that in case of destruction of intermolecular acid-labile ties of SMC it can be possible to get CDP with high molecular weights with good colloidal-chemical properties by maintaining a significant amount of the polypeptide groups in the structure of CDP.

Fermented milk compositions participating in the technology of manufacturing the products of collagen dissolution, of particular interest was the formation of the biochemical and biophysical characteristics of the finished product in the process of the influence on the raw materials of different fermentological factors, the use of simulation to identify the dominant factors responsible for the formation of quality characteristics of the product.

CDP was produced by processing raw materials in SMC after preliminary alkaline-salt treatment. As a resource material there was used the non- standard raw hide [3] processed according to the method of production of leather of chrome tanning [4]. Further processing was carried out according to the scheme presented in Fig. 1. in the environment of BPWin by applying the methodology of functional modelling IDF0.

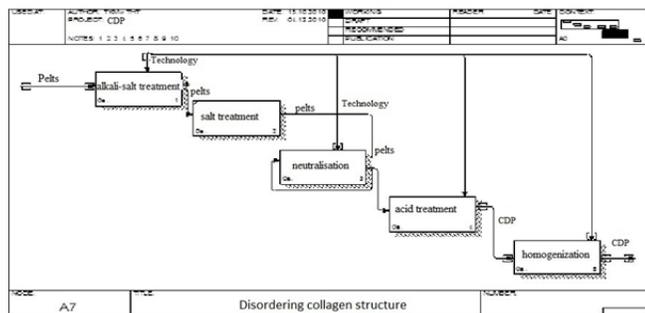


Figure 1 - Algorithm of processing raw materials to produce CDP

Acid treatment of collagen-containing raw material was carried out by SMC, which was obtained through cultivation of symbiosis of kefir grains in pasteurized cheese whey [5].

The figure shows that disordering the structure of collagen to obtain CDP is a complex and multistage process; its results depend on many factors. Given this fact and the features of object modeling, mathematical modeling has been chosen as the basic method of solving the task.

The study of the process of disordering in the structure of collagen to obtain CDP refers to the practical tasks and

requires simple mathematical apparatus. All the calculations were done in Microsoft Excel.

As input variables, there have been selected treatment temperature and titrated acidity. The processing temperature must be optimal, since its lowering can cause the inhibition of the metabolic processes of microorganisms and a temperature increase will lead to their death. The quality of the resulting product depends on the degree of destruction of acid-labile links that is provided by the action of organic acids and enzymes present in the SMC. The decrease of titrated acidity to 200°T and more no longer gives a positive result due to the insufficient quantity of lactic acid in SMC. The processing temperature significantly influences the productivity of the microorganisms, consequently, on the degree of pulping. The degree of acid-labile links destruction depends on the value of titrated acidity.

Mass fraction of ash and fatty substances have been used as output variables. Mass fraction of ash is an indicator characterizing the composition of SDP, as well as additional contaminants that may reduce the quality of the finished product.

Mass fraction of extracted by organic solvents substances can characterize the change in the composition of the product in the process of treatments due to the initiation of extraction by the action of symbiosis of microorganisms in the temperature intervals, and the initial presence of fatty substances in the starter. The pH of chlorine-potassium extract characterizes the presence of H^+ ions in the test media, which is important when carrying out acid treatment of collagen-containing raw materials. The yield of gelatin is the criterion characterizing the degree of disordering in the structure of collagen, release of low molecular weight proteins under varied conditions of the process of CDP obtaining.

At first there was conducted a preliminary experiment to determine the factor space. The results of the preliminary experiment also showed that output variables in the constraints imposed on the input variables may change according to the linear law. Given these circumstances, to construct a mathematical model there was selected the plan of complete factorial experiment. The number of experiments for two input variables ($k = 2$) is: $N = 2k = 4$.

Experiments have been performed in the laboratory. To compensate for systematic errors of the experiment there has been determined the sequence of implementation of experiments using the table of random numbers (randomization).

Statistical analysis of the significance of the estimated coefficients of the model and verification were performed in accordance with the formulas of regression analysis.

The coefficients of the regression equation were determined by the method of least squares according to the methodology [6]. Below is the augmented matrix of experimental design in coded (X_1, X_2) and natural (Z_1, Z_2) variables, and the results of the experiment to determine the mass fraction of mineral substances (Y_1), the mass fraction of fatty substances (Y_2), pH

of chlorine-potassium extract (Y3), the yield of gelatin (Y4) (table 1).

Table 1 - Augmented matrix of planning a full factorial experiment 2

X_0	X_1	X_2	X_1X_2	Z_1	Z_2	Y_1	Y_2	Y_3	Y_4
1	1	1	1	24	331	4,01	10,5	5,7	98,56
1	-1	1	-1	4	331	4,76	11,7	4,1	85,69
1	1	-1	-1	24	143	2,07	13,2	3,9	94,59
1	-1	-1	1	4	143	3,92	16,0	4,5	40,23

The mathematical model has the following form:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_{12}X_1X_2 \quad (1)$$

The coefficients of the regression equation were considered significant if the following condition was provided: $t_{calc} > t_{tbl}$, where t_{calc} - the calculated values of Student's criterion for the coefficients of the equation, and t_{tbl} is the tabulated value of Student's criterion (determined at the level equal to 0.05, the number of degrees of freedom ($f=4$)). If the calculated Fisher's criterion is less than the tabular one then the regression equations adequately describe the experimental data. The obtained models are valid only for the selected region of the input variables changes, i.e. for the temperature of the ongoing dissolution from 4 to 24°C and titrated acidity of SMC from 143 to 331°T.

To determine the optimal parameters of the process of dissolution of collagen the equation of coded variables was transformed into the equation in natural units (normalized model). They found the equation describing the dependence on the process conditions the following quality indicators CDP: gelatin melting out, mass fraction of substances being extracted with organic solvents and mineral substances, the pH of chlorine-potassium extracts.

In determining the yield of gelatin the regression equation takes the following form (2):

$$Y = 79,67 + 16,81X_1 + 12,36X_2 - 10,37X_1X_2 \quad (2)$$

The resulting equation was used to optimize the parameters of the process of dissolution of collagen. The analysis of the equation shows:

- a) the coefficient b_1 had the largest value (in absolute quantity), therefore the process temperature has the greatest influence on the yield of gelatin in the studied range;
- b) the coefficient b_{12} is significant, thus, in the investigated range of the two input variables, the joint effect of temperature and titrated acidity of the composition has a significant influence on the output variable (although in absolute quantity the value is small).

In determining the yield of gelatin after transformation the normalized equation has the following form (3):

$$Y = 4,30Z_1 + 0,30Z_2 - 0,014Z_1Z_2 - 11,50 \quad (3)$$

Fig. 2 shows a graphical representation of one of the models, which demonstrates the dependence of the studied parameters, in particular, the yield of gelatin on the parameters

of the process conducted.

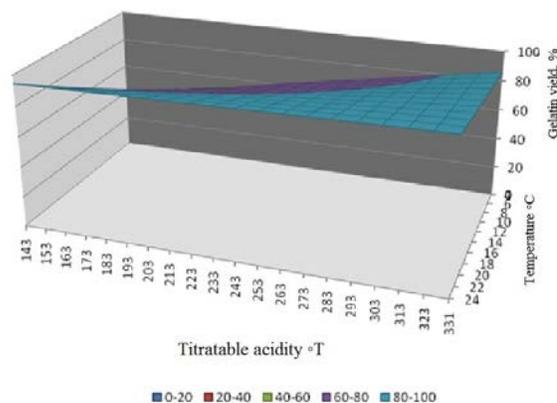


Figure 2 - Graphical representation of the dependence model of the mass fraction of the yield of gelatin upon process parameters

From figure 2 it is seen that the output change of gelatin is proportional to the temperature rise and titrated acidity. When choosing the optimal parameters of the acid dissolution of collagen in addition to gelatin melting out it is necessary to have an additional criterion by which it is possible to judge about the quality of the resulting product. By equation (3) there have been calculated theoretical values of the yield of gelatin from titrated acidity (figure 3).

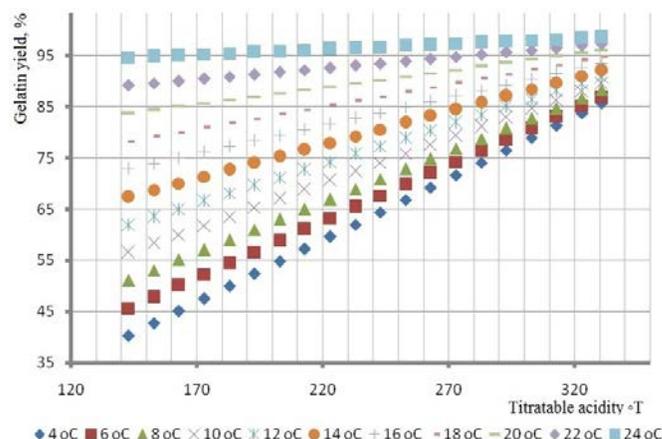


Figure 3 – Dependence of gelatin melting out in CDP on the changes in the quantity of titrated acidity

Analyzing the graphical representation of the model (figure 3), it should be noted that gelatin melting out is growing significantly with the increase in titrated acidity of the composition and temperature.

When determining the mass fraction of mineral substances, the regression equation has the following form (4):

$$Y = 3,69 - 0,61X_1 + 0,69X_2 \quad (4)$$

The analysis of the equation shows:

- a) the coefficient b_2 has the largest value (in absolute quantity), consequently, the greatest influence on the mass fraction of mineral substances in the investigated range is provided by the titrated acidity of fermented milk composition;
- b) the coefficient b_{12} turned out to be insignificant,

therefore, in the investigated range of the two input variables, the joint effect of temperature and titrated acidity of the composition has an insignificant impact on the output variable.

When determining the mass fraction of mineral substances after transformation the normalized equation has the following form (5):

$$Y = 2,85 - 0,07Z_1 + 0,01Z_2 \quad (5)$$

The resulting equation was used to optimize the parameters of the process of dissolution of collagen. Below (Fig. 4) for example, is the graphical representation of the model describing the dependence of the mass fraction of mineral substances on the parameters of the process

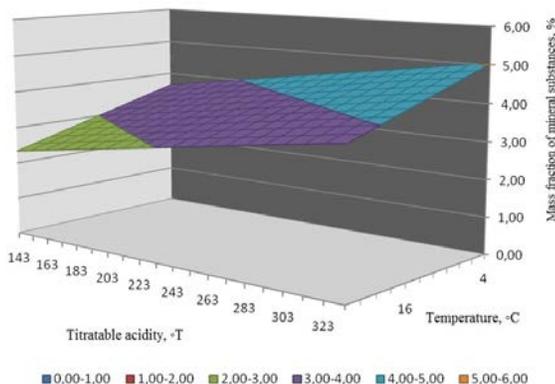


Figure 4 – Graphical representation of the dependence model of the mass fraction of mineral substances of CDP on process parameters

II. RESULTS

Thus, it was found that simulation models make it possible for computers to model and predict the processes in nonlinear complex systems, with all living systems being such ones, they are thermodynamically nonequilibrium. On the basis of the analyses of output parameters of the process of acid treatment of collagen-containing raw material, it is shown that there is no universal objective assessment indicator end of this process. It is established that a significant change in the content of proteins occur during the step of titrated acidity with 40-60°T and with the step of temperature 4°C. The optimal parameters of acid treatment of SMC are titrated acidity in the range of 300-380°T and the temperature of 20-22 degrees. It is shown that the symbiotic influence of prokaryotic organisms, as well as enzymes and acids produced by them, have the identical impact on the structure of macromolecules of collagen that allows you to set the optimal parameters of acid treatment of SMC.

On the basis of the obtained results there have been developed the theoretical basis for the symbiotic influence of prokaryotic organisms and acids produced by them on the structure of collagen:

1. The composition of the SMC, which is a symbiosis of acids, enzymes and prokaryotic organisms, which may affect the structure of collagen causing disordering.

2. The resulting product obtained in the course of the impact

of SMC on the structure of collagen is inherently a polyelectrolyte which is confirmed by its good solubility, electrical conductivity and rheological characteristics.

3. The application of SMC as acidic agent promotes the formation of high molecular weight ions (associates).

REFERENCES

- [1] Lutsenko V. A., Pinakin L. N. *Analog computers in chemistry and chemical technology*. Moscow: Chemistry. 1969. 175.
- [2] Patent № 2486258 Russian Federation MPK C14C 1/08. *A method of producing products of collagen dissolution/ Shalbuev Dm.V., Zharnikova E.V.* – № 2012100584; from 10.01.2012; publ. 27.06.2013. 13 p.
- [3] GOST 28405-90 *Raw leather. Specifications*. Moscow: Standarts' Publisher. 1990. 21 p.
- [4] *Techniques of produced by chroming leather of different thickness and assortment for the foot-wear upper and back made from hides*. CSRILL, 1983. Moscow. 185 p.
- [5] Patent № 2306345 *Method of pickling sheepskin fur raw material/ Dumnov V.S., Shalbuev Dm.V., Falileeva O.Y.* MPK C14C 1/08. publ. 20.09.2007. 9 p.
- [6] Akhnazarova A.L., Kafarov V.V. *Optimization of experiments in chemistry and chemical technology*, Moscow: Higher school. 1978. 318 p.

Building rich user profile based on intentional perspective

Sara ALAOUI^a, Younès EL BOUZEKRI EL IDRISSE^b, Rachida AJHOUN^a

^aENSIAS, Université Mohamed V, BP 713, Rabat Maroc

^bENSAK, Ecole Nationale de Sciences appliquées Kenitra

Abstract: Internet technologies evolution, from Web 1.0 to web 2.0 and web 3.0, has led us towards the definition of new requirements to be considered in the design and development of new application. It is important to note that depending solely on the request to satisfy user need is not effective. Indeed, the emergence of many researches related to the study of user behavior has enhanced the retrieval information effectiveness. Typically, the context and the user profile are the main elements to characterize the user. Hence, we aim through this contribution to build a rich profile and to provide him suitable services.

Keywords: Search Personalization, User Profiles, Retrieval information;

1. INTRODUCTION

Obviously, there is a need to have each user individualized and to get appropriate and pertinent information he is looking for. Current engine search are not very effective. Indeed, this dissatisfaction depends on several factors, including:

- The exponential number of services: the increasing of available information and services in digital format (text, audiovisual) is uncontrollable problem.
- The variety of user goals: Every user has specific context, goal and intention when searching for information [1].
- The last factor is the bad query formulation: Typically, The queries are very short, imprecise and therefore they give an incomplete specification of individual user's information needs.

In light of these challenges, several solutions and technologies are taken place to develop the retrieval search mechanism such as: adaptation contextualization personalization. We believe that these techniques are complementary. The personalization is the desired result of the adaptation

and contextualization.

Personalized systems address the previous problems by building, managing, and representing information customized for individual users. This customization may take the form of filtering out irrelevant information and/or identifying additional information of likely interest for the user. Research into personalization is ongoing in the fields of information retrieval, artificial intelligence, and data mining, among others [2]. The effectiveness of this mechanism is measured by its ability to differentiate between for example Madonna sent by a historian and the same query sent by a young looking for the updates on the famous star.

If we do a thorough reading of the previous factors and solutions we will deduct that the Knowledge about computer users is very beneficial for assisting them, predicting their future actions [3]. This was confirmed by Joel.S in her article "the weakness of traditional search technologies does not take into account the user profile [4]. The profile plays an important role in many fields and has attracted a lot of research interest. A good user profiling strategy is fundamental component in search engine personalization .User profile is a set of information allows to better understand the user needs and to predict her/his intention. The intention is a new concept that has penetrated the field of information retrieval; it is a highest degree of the comprehension of the user, it has been proposed to bridge the gap between low level, technical software-service descriptions and high level, strategic expressions of business needs for services.

As result, we will highlight in this work to the personalization approach taking into account the profile and the use requirement for predicting his future intention. This new vision is based on building a rich profile based on user intention; it is obvious

that the user profiling relies on three major pillars. These are similarity (user based, content-based), trace handling, and the prediction (machine learning, Bayesian network). We will focus in this paper to insert each choice of service by the user in his/her profile this way is more efficient because it takes into consideration the user's choice. This contribution is small part of our complete system; we propose in this paper some propositions to better understand the intent of user. The rest of this paper is organized as follows: Section (2) describes the evolution of the retrieval system. Next in section (3) we briefly present the notion of user profile and the ways of building profile. Finally in section 4, we will give a conclusion and other prospects.

1. Information retrieval system

Information retrieval (IR) is the task of representing, storing, organizing, and offering access to information items. IR is different from data retrieval, which is about finding precise data in databases with a given structure. In IR systems, the information is not structured; contained in free form in text (web pages or other documents) or in multimedia content [3]. To better understand the functioning of these systems, we propose a brief history of the research and development of information retrieval systems starting.

1.1. Classic information retrieval

Information retrieval expression was invented by (Calvin.N 1948) a student at the University of Minnesota in his master's thesis. The field of information retrieval dates back in the early 1950, shortly after the invention of computers, the pioneers of the time were keen to use the computer to automate the search. For information, which exceeded the human capacity, however, the real birth of this area was after the emergence of indexing technique. Classical information retrieval systems (IRS) considers that the user request is the sole source of knowledge about the user's information need. However, this resource is often insufficient for describing the user's preferences; it usually consists of a few short keywords, which are generally insufficient for giving a complete and accurate picture about what the user is really looking for. In fact, when the system depends on the query it returns the same result regardless of who submitted the query. Also, the same query is not essentially the same intent.



Fig 1: Classic system

The first figure shows the users' behaviors after the receipt of results, generally most user express their dissatisfaction of the returned result. Then, they will be forced to reword another query. As a result, it was necessary to rethink the construction of more efficient systems. The need for more effective information retrieval has lead to the creation of the notions of the semantic web and personalized information management,

1.2. Recent system.

In recent years, many researchers have devoted their efforts to develop retrieval search so that they make this operation an easy and a clear task. To achieve successful result in this way, it was necessary to review three primordial elements.

a. User requirement.

As users are being submerged by a large amount of information pushed them each time they request a service, there is need to have each user individualized and to get appropriate and pertinent information he is looking for. Understanding user requirements is an integral part of recent information systems. It is now widely understood that successful systems and products begin with an understanding of the needs and requirements of the users [5].the satisfaction of user needs can be achieved only when the system has a good modeling of user profiling.

b. Queries

The users tend to ask short queries, even when the information need is complex. Irrelevant Documents are retrieved as answers because on the ambiguity of the natural language (words have multiple senses).

Original request limitation:

- ✚ Ambiguity of the original query or the user typing error.
- ✚ Complexity in terms of the transformation of natural language in an understandable formal request by the used database management system.

- ✚ Complexity in terms of the formulation of requests corresponding to the responses that can be produced.

To meet this lack many web search engines today offer query reformulation suggestions by, for example, mining query logs [6].

c. System

At the system level, several approaches and techniques are given rise like, web semantic contextualization, adaptation and personalization. Importantly, all these techniques work together to offer the best services to the user. The following is an overview of the concepts of these techniques. Shilit and Theimer [7] refer to context as location, identities of nearby people and objects, and changes to these objects, the existing definition are: where you are, who you are with, and what resources are nearby the emergence of context is given. Ehu states that all Web personalization alleviates the burden of information overload by tailoring the information presented based on an individual user's needs [1]. In fact, the existing approaches answer partially the questions related to the personalization because its performance is highly related to the understudying of the behavior of user along with the ability of the system to provide different results to similar queries sent by two different users. Adaptation is a mechanism to adapt the system according to the user context. As defined Francisco J and all, Although the ultimate goal of this adaptation is always for the ultimate benefit of the end user, many approaches and techniques have been used to various degrees of experience and maturity that effectively and efficiently support context-aware adaptation[3]. The construction of user profiles is a key issue for the study of adaptation, recommendation or personalization mechanisms of information in order to take into account user's specific needs. These profiles are built and enriched according to the interactions of the user with the information system. So, if a user has low activity for example, his profile is not or badly known and the mechanisms which refer to it are less effective;

2. PROFILE

The first challenge to a personalized search system is how to collect and use the information from the user available on the web? In the Web domain, user profiling is the process of gathering information specific to each visitor, either explicitly or implicitly and representation of the user within the system.

As we have reported previously, there are two ways create user profile.

- ✓ **Implicit:** the system acquires the information necessary to build user profile in monitoring or tracing their actions. For example, if user saves a document on his computer, he is probably interested by this document or he intends to use it in the future. This way relies on the deduction of the preferences of user.
- ✓ **Explicitly feedback:** Unlike the first method this method is more significant than the previous one as it reflects the user's own real choice. In this sense, we find resources: "like" in social networks, the votes and the forms.

Despite the efforts made, the user requirement remains a challenge, and therefore, our contribution aims to help users in their research so that they would not need to modify a previous search query in hope of retrieving better results. In this paper, we will not only present a system which takes into account each interaction with the machine, but also we will register intention of user to predict the future intention.

3. INTENTION

Among definitions proposed by literature, we may find several definitions of intention.

The term intention has several different meanings. According to [8] an intention is an "optative" statement expressing a state that is expected to be reached or maintained. The notion of Intention can be seen as the goal that we want to achieve without saying how to perform it. [9] Define an intention as the goal to be achieved by performing a process presented as a sequence of intentions and strategies to the target intention. Generally the intention is presented as a composition of verb and Target and other optional parameters (giving more precision to the intention).

- Verb: determines the action allowing the determination of the meaning of intention.
- Target: represents either the object existing before the satisfaction of the intention or the result created by the action allowing the realization of the verb.
- Settings: direction, track, time.

this new concept in field of research has paved the way to the emergence of new kind of service called "intentional service"., this type of services bridge the gap between low level, technical software-service descriptions and high level language expressed by user.

4. INTENTIONAL APPROACH FOR BUILDING RICH USER PROFILE

In our user centric approach, we propose a mechanism of building a rich profile based on user intention, our objective is to achieve a heist level in user modeling to select the most appropriate service that satisfies his/her immediate intention service, and to predict to him other services . This approach will be helpful to overcome the problems already mentioned; the knowledge of user preference and the expected intention ensure a good filtering of returned result. To accomplish success result our system consists of several steps the first is the filling of the user form and select for each field the verb to complete its meaning.

1- Profile creation

Our System includes two types of users the first is a simple visitor newly entered who don't have profile; technically, this problem is referred to as cold start. It is prevalent in almost all recommender systems, and most existing approaches suffer from it [4] .the system relies on the request similarity to provide him a suitable service. The second is a loyal user who is registered and has his own profile. In this paper, we will highlight the case of the second user. This type of user has already filled out a form. Accordingly, this form contains three basic elements personal data: (name, age, address, gender), centers of interest, and finally the skills. These elements are necessary to identify and get an idea on the preferences of the user. Thus, a user ends up having their own profile. That is in the form of ontology. Using ontologies in modeling the user profile has been proposed in various applications like web search This model will allow finding similarities between the elements of the user query and the attributes of profiles; in short we have used ontology because we need semantic and clear hierarchy. Beside, the filling of form our system

propose to the user to complete each field which requires specification by choosing appropriate verb, to stay in intentional context, and to respect lexical formalism for simple intention. As already mentioned intention presented in this work uses a lexical formalism with verb, target and one or more parameters that give more specification to a verb. The verb and the target are mandatory in formulation of intention while the parameters are optional. The presence of the verb of the target allows the system to predict the user services

Fig 3: User profile form

2- Profile updating

The updating of profile comes after each submission of query the system extracts the target or product (the element searched) and the verb from the query (example: listen to the music, music is the target. In a second time, the system verifies the existence of target in the profile, in the positive case the query will be enriched by what the user chose forward.



Fig 4: Existing information

In the opposite case: the system offers to the user The system offers useful service to the user which depends on hybrid solution, user or content similarity and prediction (matching-learning and Bayesian network). We will detail these techniques later in another work. When the user chooses his appropriate service, our system automatically records this preference in their profile. In fact, the update of the profile especially the construction of a new branch requires a set of steps.

The first is searched the word in ODP ontology, Open Directory Project are emerging as an important support for ontology engineering. The system reads

from right to left, the system stops when it finds a matching between user profile and ODP element. In the following example, if the system finds that water sport exist in the profile, it must add the attribute swimming and diving; otherwise, it must build entire branch.

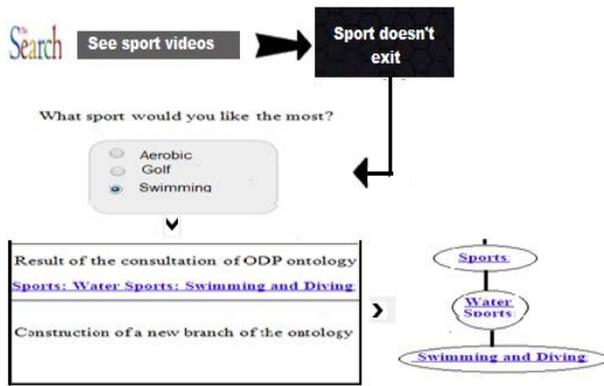


Fig 5: lack of information

4. CONCLUSION AND PROSPECTS

This work presents dynamic user profiling approach that aims to better understand user requirements, to provide him adequate and suitable service. This contribution revolves around a major technique the construction of a rich profile based on the user traces. It is necessary to highlight that this approach added important value of the profile because it is based on a clear statement of the user's preferences.

Our future works will present the implicit side for the construction of profile as similarity and deduction and we will present the weighting techniques of the profile elements.

I. REFERENCE

- [1] Aljoumaa, K., Assar, S. and Souveyet, C. Reformulating User's Queries for Intentional Services Discovery Using an Ontology-Based Approach, 4th IFIP Int. Conf on New Technologies, Mobility and Security (NTMS), Paris, France, pp. 1-4, 2011.
- [2] Susan Gauch ,Mirco Speretta, User profiles for personalized information access, The adaptative web: methods and strategies of web personalization, Springer-Verlag, Berlin, Heidelberg 2007.
- [3] J.Iglesias, P. Angelov , A.Ledezma and A.Sanchis "Creating evolving user behaviour profiles automatically", IEEE Trans, Knowl. Data eng.,2011.

[4] Petrattos,P.(2006) Information retrieval systems: A perspective on human computer interaction.Journal of Issue in informing Science and information Technology,3(1),511-518

[5] Maguire, M., & Bevan, N. User requirements analysis: a review of supporting methods. Paper presented at IFIP 17th World Computer Congress, Montreal, Canada, Aug (2002).

[6]J. Huang and E. N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In Proceedings of CIKM, pages 77–86, 2009.

[7] Schilit, B., Theimer, M. Disseminating Active Map Information to Mobile Hosts. IEEE Network, 8(5).

[8] J.Hang and E.N efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In proceeding of CIKM,pages 77-86, 2009.

[9] S.GAUCH , J.CHAFEE "Ontology based personalized search and browsing", Web Intelligence and Agent Systems, vol1 no. 3-4, 2003 [5] N. F. Noy, D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology", Stanford Knowledge Systems Laboratory Technical Report KSL-01-05, March 2001, Available at 101-noy-mcguinness.html

Augmented Reality in Radiofrequency Ablation of the Liver Tumours

Lucio Tommaso De Paolis, Francesco Ricciardi, Cosimo Luigi Manes

Department of Engineering for Innovation
University of Salento
Lecce, Italy

Abstract— Minimally Invasive Surgery is a surgery technique that provides evident advantages for the patients, but also some difficulties for the surgeons. In medicine, the Augmented Reality (AR) technology allows surgeons to have a sort of “X-ray” vision of the patient’s body and can help them during the surgical procedures. In this paper we present two applications of Augmented Reality that could be used as support for a more accurate preoperative surgical planning and also for an image-guided surgery. The AR can support the surgeon during the needle insertion for the Radiofrequency Ablation of the liver tumours in order to guide the needle and to have an accurate placement of the surgical instrument within the lesion. The augmented visualization can avoid as much as possible to destroy healthy cells of the liver.

Keywords - *Augmented Reality, medical images, minimally invasive surgery, RF ablation*

I. INTRODUCTION

The actual trend in surgery is the transition from open procedures to minimally invasive interventions. Using this surgery technique visual feedback to the surgeon is only possible through the laparoscope camera. Furthermore direct palpation of organs is not possible. Minimally Invasive Surgery (MIS), such as laparoscopy or endoscopy, has changed the way to practice the surgery. It is a promising technique and the use of this surgical approach is nowadays widely accepted and adopted as a valid alternative to classical procedures.

The use of MIS offer to surgeons the possibility of reaching the patient’s internal anatomy in a less invasive way. This reduces the surgical trauma for the patient. The diseased area is reached by means of small incisions made on the patient body called ports. Specific instruments and a camera are inserted through these ports. The surgeon uses a monitor to see what happens on the surgical field inside the patient body.

With the use of MIS the patient has a shorter hospitalizations, faster bowel function return, fewer wound-related complications and a more rapid return to normal activities. These advantages have contributed to accept these surgical procedures. If the advantages of this surgical method are evident on the patients, these techniques have some limitations for the surgeons. In some systems, the

imagery is in 2D. With this limitation the surgeon needs to develop new skills and dexterity in order to estimate the distance from the anatomical structures. We have to consider that he had to work in a very limited workspace. Modern systems use two cameras to offer a 3D view to the surgeon but this technology is useful when the working volume is not too small.

Medical images (CT or MRI) associated to the latest medical image processing techniques could provide an accurate knowledge of the patient’s anatomy and pathologies. These informations can be used to guide surgeons during the surgical procedure to improve the patient care.

The Augmented Reality (AR) technology has the potential to improve the surgeon’s visualization during the MIS surgery. This technology can “augment” surgeon’s perception of the real world with the use of information gathered from patient’s medical images.

AR technology refers to a perception of a physical real environment whose elements are merged with virtual computer-generated objects in order to create a mixed reality. The merging of virtual and real objects in an AR application has to run in real time. Virtual objects have to be aligned (registered) with real world structures. Both of these requirements guarantee that the dynamics of real world environments remain unchanged after virtual data has been added [1].

The use of AR technology in medicine makes possible to overlay virtual medical images of the organs on the real patient. This allows the surgeon to have a sort of “X-ray vision” of the patient’s internal anatomy.

In order to obtain a correct alignment of virtual and real objects a registration procedure is needed. This procedure requires the real time position tracking of the objects in the real world. This task can be accomplished with the camera, recognizing in some way the position and alignment of some objects in the scene, or with special devices called tracker. In surgery using a tracker usually improves the system uncertainty.

Using AR in surgery produces a better spatial perception and a reduction in the duration of the surgical procedure.

The aim of this paper is to present an AR system that could be used as support for a more accurate surgical preoperative planning and also for image-guided surgery.

The application can support the surgeon during the needle insertion in the radiofrequency ablation of the liver tumours. We modified the library on which our application is based to support an optical tracker that could improve the performance of the system.

II. PREVIOUS WORKS

Many research groups are now focusing on the development of systems that assists surgeons during the minimally invasive surgical procedures.

Furtado and Gersak [1] present some examples of how AR can be used to overcome the difficulties inherent to MIS in the cardiac surgery.

Samset et al. [3] present some decision support tools. These tools are based on concepts in visualization, robotics and haptics and provide tailored solutions for a range of clinical applications.

Bichlmeier et al. [4] focus on the problem of misleading perception of depth and spatial layout in medical AR and present a new method for medical in-situ visualization.

Navab et al. [5], [6], [7] present a new solution for using 3D virtual data in many AR medical applications. They introduce the concept of a laparoscopic virtual mirror, a virtual reflection plane within the live laparoscopic video.

De Paolis et al. [8] present an AR system that can guide the surgeon in the operating phase. The main goal of the system is to prevent erroneous disruption of some organs during surgical procedures. They provide distance information between the surgical tool and the organs and they use a sliding window in order to obtain a more realistic impression that the virtual organs are inside the patient's body.

Nicolau et al. [9] present a real-time superimposition of virtual models over the patient's abdomen in order to have a three dimensional view of the internal anatomy. The authors have used the developed system in an operating room and to reduce the influence of the liver breathing motion they have tried to simulate the patient's breathing cycle.

LiverPlanner [10], [11] is a virtual liver surgery planning system developed at Graz University of Technology that combines image analysis and computer graphics in order to simplify the clinical process of planning liver tumor resections. The treatment planning stage enables the surgeon to elaborate a detailed strategy for the surgical intervention and the outcome of pre-operative planning can then be used directly for the surgical intervention.

Maier-Hein et al. [12] present a system developed for computer-assisted needle placement that uses a set of fiducial needles to compensate for organ motion in real time; the purpose of this study was to assess the accuracy of the system in vivo.

Stüdeli et al. [13] present a system that provides surgeon, during placement and insertion of RFA needle, with information from pre-operative CT images and real-time tracking data.

III. AR FOR THE RF ABLATION OF THE LIVER TUMOUR

Hepatic cancer is one of the most common solid cancers in the world. Hepatocellular carcinoma (HCC) is the most common primary hepatic cancer. The liver is often the site of metastatic disease, particularly in patients with colorectal adenocarcinoma.

The use of chemotherapy for malignant form of liver cancer rarely led to good results in long-term survival rate. Surgery led to the best results for hepatic cancer care. Unfortunately only from 5 to 15 percent of patients with liver cancer can undergo to a potentially curative resection of the liver cancer [23].

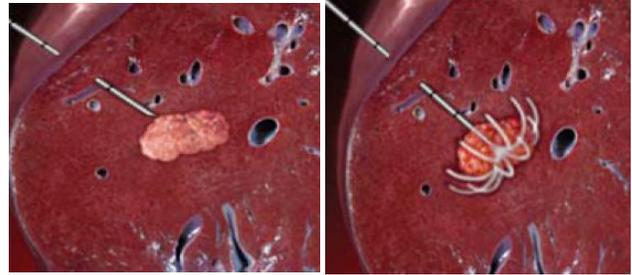


Figure 1. The effect of the RFA technique at the liver tumour

Patients with confined disease of the liver could not be candidates to resection because of multifocal disease. Another factor that could preclude surgical treatment is the proximity of tumor to key vascular or biliary structures. This preclude a margin-negative resection potentially unfavourable in case of presence of multiple liver metastases. Very often the tumor is associated to a pre-existent cirrhosis that can further reduce resection margins.

Liver transplant is the only radical therapy that eliminates the risk of recurrence. Unfortunately it can't be always used. So, since most of patients with primary or malignancies confined metastatic at the liver are not candidates for surgical resection, different approaches to control and potentially cure liver diseases were developed.

The Liver Radiofrequency Ablation (RFA) is a minimally invasive treatment for liver cancer used since 1980's. It consists in the placement of a needle inside the liver parenchyma in order to reach the centre of the tumor lesion. When the lesion center is reached, an array of electrodes is extracted from the tip of the needle and it is expanded in the tumor tissue. From these electrodes is injected in the tumor tissue a radiofrequency current that causes tumor cell necrosis for hyperthermia (the local temperature is higher than 60 °C and cancer cells are more sensitive to heat than normal cells).

In Figure 1 is shown the needle insertion and array expansion of the RFA technique on the liver tumour.

One problem in using radiofrequency tumor ablation technique is the correct placement of the needle. To ensure a maximum efficacy of the treatment the needle has to reach the center of the tumor.

Today surgeons use ultrasound, CT or MNR images acquired during the needle placement in order to correctly

direct the needle towards the center of the tumor. The use of these two-dimensional images makes the insertion procedure very difficult and requires sometimes more than one insertion.

In addition, the surgeon, in order to destroy all tumoral cells, applies the RFA on an extended area of the liver. In this way a large number of healthy cells are also destroyed. This practice can cause to the patient a number of other different consequences. To reduce this problem is of primary importance to reach the center of the tumor.

A guidance system of the needle in tumour ablation procedures can be obtained using Augmented Reality technology. With the superimposition of the virtual models of the patient's anatomy (liver, vessels, biliary ducts, cancer, etc) exactly where are the real ones, it is possible to make the needle placement task less difficult. In this way the surgeon has a sort of x-ray 3D vision of the patient internal anatomy.

The purpose of this AR application is to provide a guidance system that can help the surgeon during the needle insertion in liver RFA. The position and orientation of the ablation tool are measured using some reflective spheres attached to. These spheres are detected by an optical tracker that measure the position and orientation of the real ablation tool. These measure permits to orient the virtual ablation tool in the virtual world where virtual models of the patient anatomy are placed.

To achieve a correct augmentation it is necessary to have a perfect correspondence between the virtual organs and the real ones. This is very important in an image-guided surgery application because a very small error in the registration phase can cause a serious consequence on the patient due a bad alignment of the virtual organs on the real ones.

The tracking system is also used in order to permit the overlapping of the virtual organs on the real ones. The registration phase is one of the most delicate step in an AR system for surgery. An accurate registration is necessary to obtain a correct alignment of virtual models on the real ones. The registration is obtained using fiducial points. These points were defined on the patient body and identified on the patient tomographic images. During the registration phase the surgeon is asked to touch on the patient body with the tracker measure tool the point that the application shows on patient 3D models. In this way the application can establish the transformation between the "real" world and "virtual" world. Once registered the patient should not be moved. If the patient position can change during surgery it is necessary to fix to the patient body a tracker tool that permits to measure its position and orientation.

The novel aspect of our application is the use of an optical tracker that should decrease the guidance uncertainty in needle positioning.

IV. DEVELOPED APPLICATION

The application is provided of an user interface designed to be simple and functional at the same time. In the left side of that interface we placed the application control

panel. On the right-top window we show 3D models and the augmented reality scene. On the right-bottom there are the three smaller windows where axial, coronal and sagittal views of CT dataset are placed.

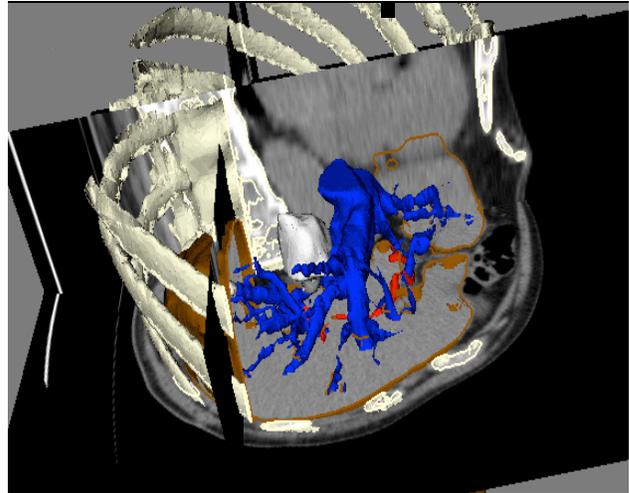


Figure 2. Clipping visualization applied to the liver and thoracic cage

The application offers to surgeon the possibility to study the case study before going in operating room. There is the possibility to apply a clipping modality that permits the surgeon to dissect the model and study its internal structure changing the opacity of the organs (Figure 2,3). The dissection could be made along the three principal axes of tomographic images.

The application features described till now are part of what we consider the pre-operative planning task. During this task the surgeon can use the application to study the pathology in a more simple and natural way than that provided by simple CT slice visualization.

For the navigation and augmentation task are devoted to the surgery room. Here the surgeon needs to use the optical tracker and to carry out the registration task. When the registration process is complete, a virtual ablation tool is shown in 3D view. It is coupled with the real ones and follows its movements.

In Figure 5 is shown the augmented visualization of the 3D virtual model over the patient's body (a dummy) during a preliminary test in the operating room. In the image is shown also the model of the needle used for RFA. This visualization should guide the surgeon during the needle insertion in the radiofrequency ablation of the liver tumour.

V. USED TECHNOLOGIES

A reconstruction of the 3D model of the anatomical structures of the patient is required in order to improve the standard slice view.

An efficient 3D reconstruction of the patient's organs is generated by applying some segmentation and classification algorithms to medical images (CT or MRI) of the patient.

The image segmentation consists in the identification of the pixels that belong to a specific anatomical structure. This process can manual, semi-automatic or fully automatic. The grey levels in the medical images are replaced by colours that are associated to the different anatomical structures [14]. After the segmentation of each slice of the dataset the software can reconstruct the 3D model using a model maker. This algorithm combines the results of segmentation of each slice and the slice tickness parameter to build a tri-dimensional model.

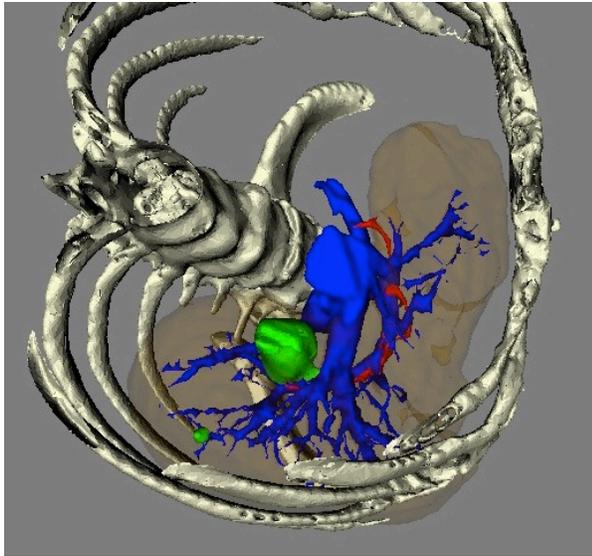


Figure 3. 3D model from the CT dataset

Nowadays there are different software used in medicine for the visualization and the analysis of medical images and the 3D modelling of human organs. Mimics [15], 3D Slicer [16], and OsiriX [17] play an important role among these

tools. Some of them are also open-source software.

In our application we have used 3D Slicer for the building of the 3D model of the patient’s organ. 3D Slicer is a multi-platform open-source software package for visualization and image analysis. It has many builtin segmentation algorithms and also a model maker module. Fig. 1 shows a 3D model built from CT scan of the patient with the liver tumors models in green.

The AR guidance application was developed using Image-Guided Software Toolkit (IGSTK) framework [18]. It is a set of high-level components integrated with low-level open source software libraries and application programming interfaces.

In the developed AR platform we use an optical tracker. This tool is able to detect some retro-reflective spheres intentionally introduced in the surgical scene. These spheres are placed on the surgical tools. They provide within a defined coordinate system the real-time spatial measurements of the location and orientation of the surgical instruments used during the surgical procedure.

In the first prototype of the system we have used the Polaris Vicra optical tracker [19]. After the test in the operating room we changed the tracking system to ensure surgeons more freedom of movement. For this reason we decided to use the Bonita Vicon tracking system [20]. This tracking system uses a variable number of separated cameras that can be better positioned in the operating room.

A specific library has been developed and tested for interfacing with the Bonita Vicon tracker. This library was integrated inside IGSTK framework.

At the base of the functioning of IGSTK there is the use of a state machine that allows to increase the safety and robustness of the toolkit [21].

The use of a state machine, in fact, allows to limit and control the possible behaviors of application in order to ensure that it is always in a formal state planned in the

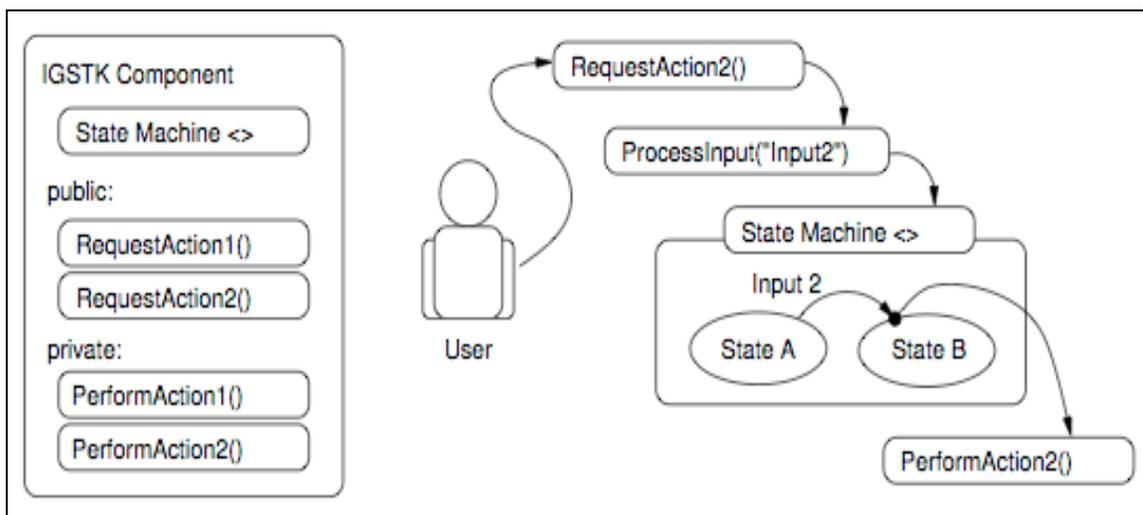


Figure 4. Working modality of IGSTK library components

design phase. This guarantees a reproducible and deterministic behavior that eliminates the risk of application design formal errors.

The separation between public and private interface of a IGSTK component (object) is shown in Figure 4.

When a user send a request for an action through a call of a public method, this request is translated into an input for the state machine. The state machine taking into account the current state and the sent input changes its state as expected in the design phase. In any case, the object will always be in a known state because each type of behavior has been programmed.

The IGSTK Tracker component communicates with the tracker to obtain the position and orientation relative to each tool that is present in the acquisition volume. This information is then transmitted to other IGSTK components who request it.

The two classes of IGSTK framework that provide for trackers management are:

- “igstk::Trackers” that serves to manage the status information of the tracking system;
- “igstk::TrackerTool” that manages the information associated with each tool present on the scene.

The “igstk::Trackers” class uses the interface of the “igstk::Communication” class for the management of communications. This allows to prevent blockage of the application in the case of any lack of connection with the tracker. In this case an error event is generated in a non-blocking mode. The tracker acquired data are stored into a memory buffer.

The class “igstk::PulseGenerator” generates the clock used by the application in order to synchronize all events. In this way, reading and updating of the data are asynchronous operations and this allow high standards of safety and performance.

To calibrate the instrument used in the surgical application and obtain the transformation elapsing between the tool and the point of interest has been implemented the PivotCalibration algorithm. This algorithm permits to store the position and orientation of the tracker tool holding fixed the position of the instrument tip and moving the tool in the space [22].

VI. APPLICATION TESTING

The first test of application was made on a dummy. We tried to overlap the 3D models of a real patient on the dummy, as shown in Figure 5. These models were generated from a sample, anonymized CT dataset. This test was not successful because there was a misalignment between the dummy and the virtual models. This is due to a different thoracic girth between patient and dummy.

To overcome this problem we decided to make a preliminary qualitative test of application in the operating theater. This test is designed to evaluate the application uncertainty in the operating room and also all the possible issues related to a live use in the operating room.

A selected case of patient with superficial liver cancer

lesions was found by surgeons. The patient will undergo to a traditional open surgery liver RFA.

After a traditional needle insertion we started the application and made registration. At this moment it was verified if a good overlap with tumor and real organs was obtained. In this case we obtained good qualitative result.

The test doesn't add risks for the patient, if compared with traditional RFA surgery, because the application was not used to guide the needle insertion. Anyway the patient was informed on the nature of the experiment and signed an informed consent form.

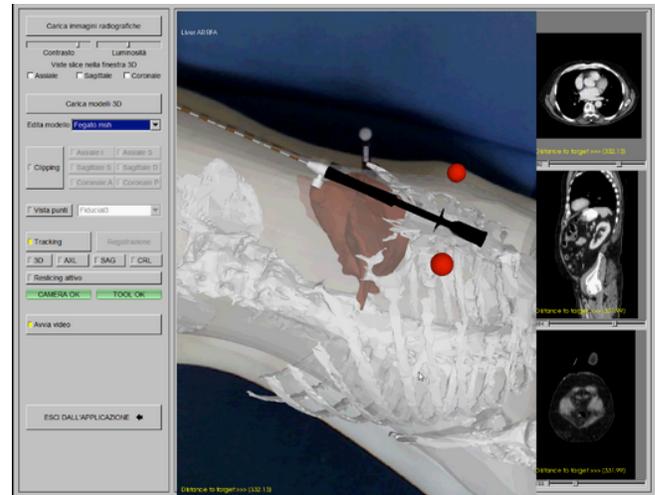


Figure 5. Augmented visualization of the 3D virtual model over the patient's body (a dummy)

VII. CONCLUSIONS AND FUTURE WORK

In this paper we present a guidance system for needle placement in radiofrequency ablation of the liver tumour.

The application has been tested previously on a dummy and afterwards a first test has been carried out in the operating room during an open surgery procedure for the liver tumour resection.

The laboratory test consists in a simple overlapping of real patient models obtained from a simple anonymized CT dataset on a dummy. The results of this test were not as expected because there was a significantly difference in thoracic girth between patient and dummy.

We decided to move to the operating room to test the application in a qualitative manner on a real patient. The models were built from patient's CT scans. The test in the operating room has also the aim to evaluate all the possible issues of that operating scenario. In Figure 6 is shown a phase of the first test in operating room.

The test on the real patient led to good results. A correct overlap of virtual models on real patient was obtained in this case.

The test in the operating room produced interesting results for all the aspects related to the constraints of the

particular environment. Operating rooms are overloaded of systems and devices and an efficient use of space is mandatory. In this first test the correct position of the devices close to the operating table has been defined and a precise definition of the more appropriate fiducial points used for the registration phase has been decided.

Surgeons judged the tracker volume of interaction to be not sufficient. For this reason we decided to change the model of the optical tracker used in this first test to obtain a bigger volume of interaction. A specific library was developed and tested for interfacing with this new tracker. This library was integrated within the IGSTK framework. This new tracker should be tested in the laboratory to evaluate its measure uncertainty and then in the operating room to evaluate if the new volume of interaction can be sufficient for the surgeons.

Anyway surgeons were excited and think that the AR technology can help them in the needle placement task.



Figure 6. First test in the operating room

The next test will be done on a pig liver in order to measure the precision of the image-guided application. We're planning also to design a quantitative test in operating room to evaluate what is the application guidance uncertainty.

We are also taking into account the possibility to include in the system the simulation of the virtual model deformations due to the breathing of the patient.

REFERENCES

- [1] S. Maad, "Augmented Reality. The Horizon of Virtual And Augmented Reality: The Reality of the Global Digital Age", Intech, January 2010, ISBN 978-953-7619-69-5.
- [2] H. Furtado and B. Gersak, "Minimally Invasive Surgery and Augmented Reality. New Technology Frontiers in Minimally Invasive Therapies", 2007, pp. 195-201.
- [3] E. Samset, D. Schmalstieg, J. Vander Sloten, A. Freudenthal, J. Declerck, S. Casciaro, Ø. Rideng, and B. Gersak, "Augmented Reality in Surgical Procedures", SPIE Human Vision and Electronic Imaging XIII, 2008.
- [4] C. Bichlmeier, F. Wimmer, H. S. Michael, and N. Nassir, "Contextual Anatomic Mimesis: Hybrid In-Situ Visualization Method for Improving Multi-Sensory Depth Perception in Medical Augmented Reality", Proc. Sixth IEEE and ACM Int. Symposium on Mixed and Augmented Reality (ISMAR '07), Nara, Japan, 2007, pp. 129-138.
- [5] N. Navab, M. Feuerstein, and C. Bichlmeier, "Laparoscopic Virtual Mirror - New Interaction Paradigm for Monitor Based Augmented Reality", Proc. IEEE Virtual Reality Conf. 2007 (VR 2007), Charlotte, North Carolina, USA, 2007, pp. 10-14.
- [6] C. Bichlmeier, S. M. Heining, M. Rustae, and N. Navab, "Laparoscopic Virtual Mirror for Understanding Vessel Structure: Evaluation Study by Twelve Surgeons", Proc. 6th IEEE International Symposium on Mixed and Augmented Reality, Nara, Japan, 2007, pp. 1-4.
- [7] C. Bichlmeier, F. Wimmer, S. M. Heining, and N. Navab, "Contextual Anatomic Mimesis: Hybrid In-Situ Visualization Method for Improving Multi-Sensory Depth Perception in Medical Augmented Reality", IEEE Proc. Int. Symposium on Mixed and Augmented Reality, Nara, Japan, 2007.
- [8] L. T. De Paolis, M. Pulimeno, M. Lapresa, A. Perrone, and G. Aloisio, "Advanced Visualization System Based on Distance Measurement for an Accurate Laparoscopy Surgery", Proc. Joint Virtual Reality Conf. of EGVE - ICAT - EuroVR, Lyon, France, 2009, pp. 17-18.
- [9] S. Nicolau, A. Garcia, X. Pennec, L. Soler, X. Buy, A. Gangi, N. Ayache, and J. Marescaux, "An augmented reality system for liver thermal ablation: Design and evaluation on clinical cases", Elsevier, 2009.
- [10] LiverPlanner, <http://liverplanner.icg.tu-graz.ac.at>
- [11] B. Reitingner, A. Bornik, R. Beichel, G. Werkgartner, and E. Sorantin, "Tools for augmented reality based liver resection planning", Proceedings of the SPIE Medical Imaging 2004: Visualization, Image-Guided Procedures, and Display, pages 88-99, San Diego, February 2004.
- [12] L. Maier-Hein, A. Tekbas, A. Seitel, et al., "In vivo accuracy assessment of a needlebased navigation system for CT-guided radiofrequency ablation of the liver", Medical Physics, 2008, Vol. 35, No. 12, 5385-5396, 0094-2405.
- [13] T. Stüdeli, D. Kalkofen, P. Risholm, et al., "Visualization tool for improved accuracy in needle placement during percutaneous radiofrequency ablation of liver tumors", Medical Imaging 2008: Visualization, Image-Guided Procedures, and Modeling, Pts 1 and 2, Vol. 6918, B9180-B9180, 0277-786X.
- [14] T. S. Yoo, "Insight into Images: Principles and Practice for Segmentation, Registration, and Image Analysis", A K Peters, Ltd, 2004.
- [15] Mimics Medical Imaging Software, Materialise Group, <http://www.materialise.com/>
- [16] 3D Slicer, <http://www.slicer.org>
- [17] OsiriX Imaging Software, <http://www.osirix-viewer.com>
- [18] K. Clearya, L. Ibanez, and S. Ranjan, "IGSTK: a software toolkit for image-guided surgery applications", Conference on Computer Aided Radiology and Surgery, Chicago, USA, 2004.
- [19] NDI Polaris Vicra, <http://www.ndigital.com>
- [20] Vicon Bonita, <http://www.vicon.com/products/bonita.html>
- [21] Kevin Clearya, Patrick Cheng, Andinet Enquobahrie, Ziv Yaniv, "IGSTK: The Book", May 29, 2009
- [22] A. Lorsakul, J. Suthakorn and C. Sinthanayothin, "Point-Cloud-to-Point-Cloud Technique on Tool Calibration for Dental Implant Surgical Path Tracking", Proc. SPIE 6918, Medical Imaging 2008: Visualization, Image-guided Procedures, and Modeling, March 17, 2008
- [23] D. Nagorney, J. Van Heerden, D. Ilstrup, et al., "Primary hepatic malignancy: surgical management and determinants of survival", Surgery, 1989, vol. 106, pp. 740-748.

Management of intangible assets within health care industry. A comparative study between Sweden and Poland

Dorota Jelonek, Czestochowa University of Technology, jelonek@zim.pcz.pl
Amra Halilovic, County Council of Dalarna, amra.halilovic@ltdalarna.se

Abstract - In the sector of medical services the human capital has a significant meaning among intangible resources, its knowledge and competences, relational internal resources with patients and structural resources. The paper presents selected case studies of management of intangible assets within health care industry in Sweden and in Poland. It was indicated that the informational technology, especially the internet technology is an effective support in management of intangible assets within health care industry

Keywords— Health care industry, Intangible assets, National quality registries,

I. INTRODUCTION

An organization's strategy describes how it intends to create value for its stakeholders (customers, shareholders, citizens etc) by exploiting their internal resources and capabilities [1, 2].

Traditionally, these strategies were focused on competitive forces and physical resources, such as machines, equipments and financial capital. In today's economy, the businesses has shifted from seller's markets to buyer's markets and customers have become more demanding. Factors like innovation, speed, differentiation have become critical [3, 4].

"More recently, intangible assets have been identified as key resources and sources of competitive advantage...It is argued that a sustainable competitive advantage results from the possession of resources that are inimitable, not substitutable, tacit in nature, and synergistic." [4]

II. INTANGIBLE ASSETS

There is the growing gap between the book value and trading value of organizations is the reason of greater interest in the conception of intellectual capital. This intellectual capital is identified by many researchers as intangible assets which can be used for building the value of the organization. [5]

International Accounting Standards Board defines an intangible asset as: *"an identifiable non-monetary asset without physical substance."* [6] There are two major characteristics of intangible assets: they neither create value nor generate growth and they represent capabilities and potential for future growth and income. [7].

Table 1 below describes commonly accepted categories of intangible assets: human assets, relational assets and structural assets. [4].

Table 1: Taxonomy and short description of intangible assets

INTANGIBLE ASSETS	DESCRIPTION
HUMAN ASSETS	Include: Skills, competence, commitment, motivation, loyalty of employees. Key components: know-how, technical expertise, problem solving capability, creativity, education, attitude, and entrepreneurial spirit.
RELATIONAL ASSETS	Include: all forms of relationships a company has with its stakeholders and customers. Key components: licensing agreements, partnering agreements, contracts and brand image.
STRUCTURAL ASSETS	Include: all intangibles that stay with the organisation such as corporate culture, routines and practices such as tacit rules, intellectual property, processes whose ownership is granted to the company by the law.

It is worth indicating that among identified about 200 intangible resources by representatives of the resource school, only the part of them is an object of turnover in the market [8].

Common health care industry intangible assets are mostly focused on relational assets as: patient relationships; medical, dental, and other professional licenses; certificates of need; facility operating licenses and permits; physician (and other professional) employment agreements; physician (and other professional) noncompetition agreements; executive (and other administrator) employment agreements; executive (and other administrator) noncompetition agreements; administrative services agreements; medical (and other professional) services agreements; facility or function management agreements; equipment use or license agreements; equipment and other supplier purchase agreements; joint venture agreements; joint development or promotion agreements. In human and structural assets counts: patent files and records (manual and electronic); electronic medical records computer software; medical and administrative assembled workforce; office systems, procedures, and manuals; position or "station" procedures and manuals; service marks and service names; a professional's personal goodwill; an entity's institutional goodwill; medical (other professional) staff privileges. [9].

Most health care industry participants own and operate intangible assets. These intangible assets can be industry-specific (e.g., patient charts and records, certificates of need, professional and other licenses), or they can be general commercial intangible assets (e.g., trademarks, systems and procedures, an assembled workforce) [10].

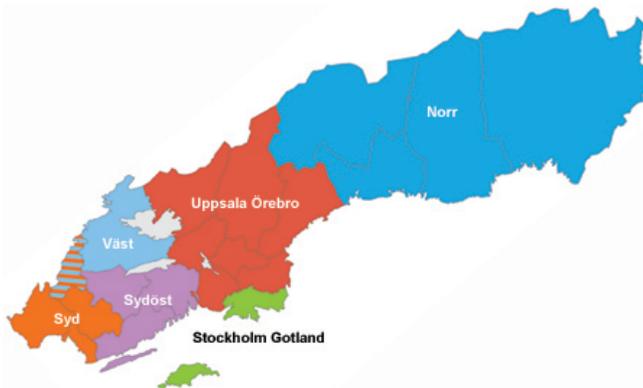
III. NATIONAL QUALITY REGISTRIES AS INTANGIBLE ASSETS IN SWEDISH HEALTH CARE INDUSTRY

National quality registries (NQR) in Sweden have a long history. From the beginning, they were created by the individuals who themselves would benefit from them in their professional lives.

NQRs are a system of quality tools which aim to continuously improve and provide good health care. This means that healthcare has to be consistent (equal treatment for all patients, throughout the country) and to ensure that treatments are fact based. NQRs contain individualized data concerning patient problems, medical interventions, and outcomes after treatment; within all healthcare production in the country. The individualized data is protected by several laws. NQRs also enable:

- Monitoring the progress made in health care, both for the individual patients, as well as at the aggregate group level (for example, a group of cancer patients).
- Following patient outcomes per county, hospital or clinic.
- Support the health care work (for example, checklists).
- Compare healthcare units' own work over time and identify areas for improvement.
- Research based on data from Sweden's healthcare units.

In Sweden there are six Regional Cancer Centres and six Competence Centres. Registers in cancer field (approximately twenty national quality registers) are organized in six Regional Cancer Centres (RCCs): Norr, Stockholm Gotland, Syd, Sydöst, Uppsala Örebro och Väst (see Figur 1 below).



Figur 1: Six Regional Cancer Centres in Sweden (source: SKL)

RCCs work for a more patient-focused, equitable and effective cancer care. These centres receive, encode, record and verify the information annually forwarded from the region to the Cancer Registry at the National Board.

RCCs have a common goal and full autonomy. There are ten criteria that specify the frame and focus of the RCC activities and its organization:

1. Design and implement a plan for the region's on prevention and early detection of cancer.

2. Manage and coordinate the region work in order to make cancer care chain more effective.

3. Have a plan that ensures cancer patients' access to psychological support, rehabilitation and good quality palliative care across the region.

4. Strengthen patients' position in their cancer care.

5. Design and implement a plan for the development of the region's cancer care.

6. Reinforce progress towards knowledge-driven cancer care.

7. Strengthen clinical cancer research both in the region and in the country.

8. Have a clear management structure with strong roots within the county, interact with other RCCs and have systems for monitoring cancer care quality.

9. Develop a strategic development plan for cancer care in the region.

10. Develop a plan for cancer care level structuring and support the implementation of the plan.

Registers in cancer field have a common technology platform, owned by county councils / regions and the management and development of RCC.

Six competence centers ("registercentrum") for the other NQRs (approx. 60) have been established [8]: Registercentrum Norr (RCN), Uppsala Clinical Research Center (UCR), QRC Stockholm, Registercentrum Västra Göteland (RVG), Registercentrum SydÖst (RCSO) and Registercentrum Syd.

In these competence centres, several registries share the costs of staff and systems that a single registry could not bear, e.g., in technical operations, analytical work, use of registry data to support clinical quality improvement, and helping to make registry data beneficial for different users. These six Competence Centres have not a common technology platform. Hence, a continued development of the registries can be assured, although the system follows a decentralized model, i.e. each register is governed by an executive board.

Results from NQRs are available to medical units and county management over the Internet. Results from NQRs are also accessible by reports like "Open Comparison and Assessment". These reports are freely available to all, including citizens.

Citizens do not have access to the results over Internet. Recently, there are some attempts (by some NQRs) to enable reports even to citizens.

IV. EXAMPLE FROM COUNTY COUNCIL OF DALARNA (SWEDEN)

The Swedish Association for Diabetology (SFD) established The Swedish National Diabetes Register (NDR) in 1996 in response to the St. Vincent Declaration, whose purpose was to persuade European countries to reduce the prevalence of diabetes complications. [11].

The NDR is maintained by the Swedish Society for Diabetology on behalf, and with the financial support, and the Swedish Association of Local Authorities and Regions. [11].

The NDR is the largest diabetes register in the world.

Approximately 90% of all Swedish people with diabetes were entered in the register in 2013. The NDR has engaged the participation of both hospitals and primary care clinics. The overall objective is to reduce morbidity and mortality, as well as to maximise the cost-effectiveness of diabetes care. The register offers a unique opportunity to monitor the quality of care in terms of risk factors and the potential complications of diabetes, as well as the evolution of treatment methods. [11].

Electronic data entry at www.ndr.nu provides a clinic with immediate access to its results, as well as county-by-county and nationwide comparison statistics. A patient's data may be entered repeatedly throughout the course of a year. Approximately 86% of NDR users sign on with personnel cards. [11].

County council of Dalarna transmits data directly from medical record to the NDR database. 100% of all county's people with diabetes were entered in the register in 2014 (by the law the patient has the right to say no to registration). All data entries are automated and validated on a continual basis.

Each specialist clinics as well as primary care clinics has immediate access to its own outcomes, as well as nationwide comparison figures. The results are based on input data for the period of time that the user selects. A diabetes care unit can autonomously generate their annual report, including nationwide comparison figures. The reports serve as a tool for monitoring and improvement efforts.

This register data is documented evidence and leads to better outcomes for patients. The critical factors for success are measuring results integrated into the overall diabetes care process, as well as training the entire team to participate in the improvement effort. Another factor that is crucial is the commitment of physicians and other professionals to measuring results, collecting data and discussing what they have learned.

The NDR in County council of Dalarna is the "success" example for management of human and structural assets in health care industry.

V. „QUICK ONCOLOGICAL THERAPY“ AS AN EXAMPLE OF MANAGEMENT OF INTANGIBLE ASSETS IN POLISH HEALTH CARE INDUSTRY

The Polish healthcare system for a dozen or so years is facing profound reforms. The most important determinants of changes are:

- limited access to Healthcare,
- the long time of waiting for a visit at the specialist
- requirements of the European Union towards Poland as the member state,
- limited funding of medical services,
- demographic changes of societies
- lack of competition among insurers,
- unequal status of public versus private healthcare providers,
- indebtedness of public healthcare institutions
- development of information technologies which facilitate

implementation of e-health system.

The basic direction of reforms in Polish healthcare system concerns implementation of e-health system. E-health system can improve prevention of illness, delivery of treatment, and support a shift from hospital care to primary care. E-Health can help to provide better citizen-centred care as well as lowering costs and supporting interoperability across national boundaries, facilitating patient mobility and safety [13]. eHealth can benefit citizens, patients, health and care professionals but and health organizations. Information technologies support management of intangible assets within health care industry.

Among many elements of intangible assets the greatest attention is concentrated on: development of internal human resources and their competences (doctors, nurses) and on external (patients) resources. The more and more greater meaning has a development of internal relations and relations with patients. In Poland like in Sweden are supported solutions which take into account formation of competence centres, particularly concerning cancers. Centres of competences assure possibility of wide consultations, supports from doctors, supports of patients, assurance of the wide access to the full information e.g. about the case history, about new methods of treatment, about facilities which have at their disposal the latest equipment, about new medicines and therapies etc.

An example of such a solution can be a program „Quick oncological therapy”, which became initiated from 1 January 2015 for patients with the suspected cancer.

Benefits which will be brought from implementation of „Quick oncological therapy" are [14]:

- shortening of queues for patients with the suspected neoplasm,
- the arrangement of the diagnostics process and treatment of patient,
- introduction of quick diagnostics and complex treatment,
- diminution of the mortality of oncological patients,
- decrease of medical costs, thanks to the detection of the illness in the early stage.

The process of „Quick oncological therapy” was presented in figure 2.

At the first stage of the process the patient with the suspicion of cancer is registered in the IT system „service of diagnostics card and the oncological treatment DiLO”. This DiLO card called also „ a green card” is a symbol of this program and assures patients quicker and more effective path of diagnostics and the oncological treatment.

In the relational capital management of hospitals can be used patient relationship management systems based on commercial CRM systems[15].

Management of intangible within health care industry in Sweden and Poland was presented on the examples of diagnostics programs and oncological treatment. The common feature of analyzed examples is the aspiration to the construction of common information and knowledge bases integrated in the scale of all the country's IT systems. Important is also to ensure the access to centres of competences development for doctors what is most important in building of the confidence of patients to hospitals and strengthenings relational capital with the patient, what in turn will contribute to the consolidation of the competitive position of the hospital in the market of healthcare services

REFERENCES

- [1] Penrose, E. T. (1959). *"The theory of the growth of the firm"*. New York: John Wiley.
- [2] Kaplan, R. S. and Norton, D. P. (2004). "The strategy map: guide to aligning intangible assets". *Strategy & Leadership*, Vol. 32 Iss 5, pp. 10-17.
- [3] Marr, B. and Spender, J. C. (2004). "Measuring knowledge assets – implications of the knowledge economy for performance measurement", *Measuring Business Excellence*, Vol. 8 No. 1, pp. 18-27.
- [4] Marr, B. (2005). "Strategic management of intangible value drivers". *Handbook of Business Strategy*, Vol. 6 Iss 1, pp. 147-154.
- [5] Jelonek, D., Chulski, A. (2014). "Technological context of health care entity intangible assets management", *Online Journal of Applied Knowledge Management*, Vol. 2, Iss 2.
- [6] IAS 38 — Intangible Assets. Available at: <http://www.iasplus.com> (20.04.2015)
- [7] Lev, B., Daum, J. (2004). "The dominance of intangible assets: consequences for enterprise management and corporate reporting" *Measuring Business Excellence*, Vol. 8 No 1, pp. 6–17.
- [8] M. Rzemieniak, The management of intangible assets of enterprises (Zarządzanie niematerialnymi wartościami przedsiębiorstw), Dom Organizatora TNOiK, Torun 2013.
- [9] Reilly, R. F. (2012). "Cost Approach of Health Care Entity Intangible Asset Valuation" *Journal of Health Care Finance*, Winter, Vol. 39 No 2, pp 1–36.
- [10] Reilly R. F. (2010). "Intangible asset valuation, damages, and transfer price analyses in the health care industry", *J Health Care Finance*, Spring, Vol. 36(3), pp.24-33
- [11] Swedish Association of Local Authorities and Regions, <http://www.skl.se> (20.04.2015)
- [12] The Swedish National Diabetes Register, <http://www.ndr.nu> (21.04.2015)
- [13] COM(2004) 356 relating to an action plan for a European e-Health area
- [14] <http://pakietonkologiczny.gov.pl/o-terapii/> (23.04.2015)
- [15] Jelonek D., Chluski A., (2010), "The possibility of using CRM systems in health care" in. Information technology in public administration and healthcare. (Możliwości wykorzystania systemów CRM w zakładach opieki zdrowotnej, in. Technologie informatyczne w administracji publicznej i służbie zdrowia). Edited by J. Goliński, A. Kobyliński, A. Sobczak, SGH, pp.35-47.

Automatic acquisition, processing and analysis of data system, using the AHP multi-criteria method

Sorin Borza, Carmen Simion,

¹**Abstract**—In this paper, an automatic system of taking over the data of the environment and their automatic transformation in GIS Data, for some intelligent map's generation, is presented. Also, the system allows, by using the multi-criteria method AHP, an objective analysis of the factors that will be taken into account during the system. These factors will be taken over automatically by using the LabVIEW Software and with the help of the acquisition board that will be connected to the computer. By using the same LabVIEW software, the factors will be memorised in an ACCESS type database. The database will be connected to the Geomedia Professional Software, with the help of which an intelligent map will be generated automatically. The system allows, not only the automatic collection of data, but also their memorization and the generation of GIS elements and an objective analysis of the collecting points of data could provide a concrete answer regarding the most and least polluted points. As a novelty element, the paper allows the analysis of the polluted factors using the multi-criteria AHP method in such automatic taking-over storage and data generation system. Also, in this paper, an economic analysis of the used system will be made. In the first part of the paper, the general aspects of the paper will be presented. In the second part, the object oriented technics of the realization of the virtual apparatus and of the data transfer in the Access database. In the third part of the paper, the AHP method will be presented. The fourth part will present the physical realization of the module system of automatic taking over data, the analysis of the polluting factors taking account the AHP method, the automatic generation of the object-oriented classes needed for the elaboration of the intelligent maps.

Keywords— databases, virtual instrument, object oriented, GIS.

I. INTRODUCTION

The development of human society has led to a negative anthropic and technogenic impact on the air quality, resulting in a significant series of adverse effects on human health, flora, fauna and ecosystems in general. In urban areas, the environmental dimension, the economic and the social one interfere the most [7].

F. A. Sorin Borza *Faculty of Engineering, Sibiu University "Lucian Blaga", 10, Victoriei Bd, Sibiu, 550024, România* (e-mail: sorin.borza@ulbsibiu.ro).

S. B. Carmen Simion . *Faculty of Engineering, Sibiu University "Lucian Blaga", 10, Victoriei Bd, Sibiu, 550024, România* (e-mail: carmen.simion@ulbsibiu.ro).

Four out of five European citizens live in urban areas and their quality of life is directly influenced by the state of the urban environment. A high quality of the urban area also contributes to reaching the priority of the revised Lisbon strategy, which is „to make Europe a more attractive place to live and invest”. The attractiveness of the European cities will enhance growth potential and also job generation, and therefore cities will be key factors for the implementation of the Lisbon Agenda. In the case of urban habitats, transportation is “guilty” for more than 50% of NO_x, particles (PM) and volatile organic compounds (VOCs), over 25% of CO₂ emissions and 80% of the noise level.

The U.S. Environmental Protection Agency's (EPA) Clean Power Plan, proposed in June 2014, would limit carbon pollution from existing power plants [8].

Vehicular air pollution has been recognized as a major anthropogenic activity responsible for deteriorating air quality in urban areas. One of the major traffic related air pollutants which have serious health effects is carbon monoxide (CO). One of the major components of traffic related air pollution is carbon monoxide (CO) that results from the incomplete combustion of fuel. CO does not participate in secondary atmospheric reactions but has approximately 210 times more affinity for the hemoglobin (Hb) than oxygen (O₂). Carbon monoxide enters the bloodstream through lungs and form Carboxy hemoglobin (COHb), a compound that inhibit blood capacity to carry oxygen to organs and tissues. Though studies have ruled out any possibility of detectable increase of genetic damage in blood cells due to moderate air pollution levels [9].

The automatic taking over and the analysis of the data using the AHP method (Analytic Hierarchy Process) and of generating the element's classes necessary to elaborate the intelligent maps consists of hardware components: sensors of the acquisition board, apparatus used for measuring the polluting factors and also, of software elements, like: LabVIEW, Geomedia Professional, Access Database, figure 1.

The data will be taken over, with the help of the sensors or with the help of the specific apparatus which will be connected to the computer. In the case in which the data will be taken over by the help of the sensors, the IEPE (Integrated Electronic Piezoelectric) sensors were used, the acquisition board NI 9234, NI Mseries. In the case in which the taken over data by the specialized apparatus, the Drager Pac III apparatus has been used for the measurement of the monoxide carbon

and the sound level meter EXTECH 407780 for the sound measurement. The Drager Pac III apparatus is equipped with a sensor for the concentration measurement of the monoxide carbon. The values showed on the display apparatus are represented in parts on a million (ppm), these values must be multiplied with 1,16, and the resulted value represented in mg/m³. The virtual apparatus will be made using the compute, by the help of the LabVIEW software. The multi-criteria analysis will be realized by a virtual apparatus. Another part (branch) of the system will allow the data memorization in the ACCESS database for their following use in the class generation process (features classes) used for the elaboration of the intelligent maps.

II. OBJECT ORIENTED PROGRAMMING IN LABVIEW

In object oriented programming with LabVIEW [5], a class consists merely of a user defined data type together with methods that can be applied to values of that data type. Once could say that object oriented programming in Labview allows the developer to create object oriented wires. Object orientation in Labview means the following:

1. Simple Inheritance. Neither multiple inheritance nor interfaces as in Java.
2. Strict encapsulation. Data of a class are always private. Public or protected data do not exist.

As in other object oriented languages, a derived class may overload an abstract method of its base class. However, the override method must have exactly the same input and output parameters as the respective method of the parent class.

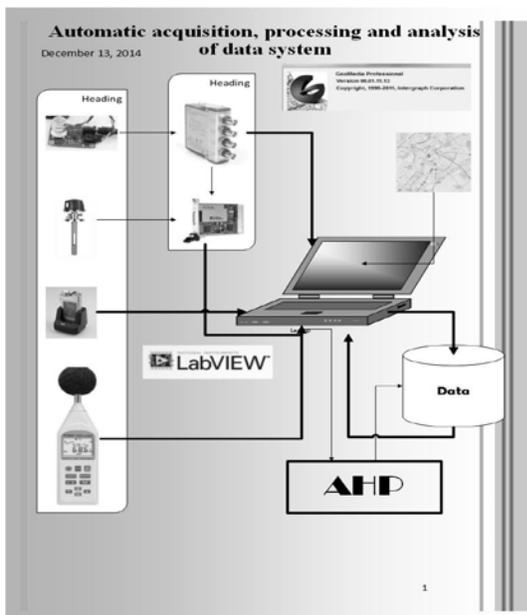


Fig. 1 The System presented in this paper

In object oriented programming for Labview [6] are three fundamental consequences.

1. Objects contain only data and no active code. Agents do not exist.

2. Labview does not have classical variables. For the same reason, in Labview has no equivalence to the concept of a *constructor* and a *destructor*. There are neither constructors nor destructors.

3. Objects can only be accessed "by value" and never "by reference".

The fundamental differences between object programming in Labview and conventional object oriented languages prevent a straightforward *implementation* of design patterns that are based on the idea of objects as entities. However, many of those design patterns are useful for designing control systems.

III. ANALYTICAL HIERARCHY PROCESS (AHP) METHOD

Analytic Hierarchy Process (AHP) is used for decision making when a decision (choice of some of the available alternatives, or their ranking) is based on several attributes that represent criteria [1]. Solving complex decision problems using AHP method is based on their decomposition in a hierarchical structure whose elements are goal (objective), criteria (sub-criteria) and alternatives. An important component of the AHP method is a mathematical model by which priorities of elements are calculated (weighted), for elements that are on the same level hierarchical structure. AHP was successfully used in environmental impact assessment for determining of weights for impact categories in paper [3]. In paper [4] AHP was used for verification of results gained by quantification of environmental aspects and impacts.

Summary of AHP method consists of converting subjective assessments to the relative importance of the criteria scores and weights. The method, developed by Saaty [1] proved to be the most common form of multi-criteria analysis. AHP input data are answers to questions such as "How important is criterion A relative to criterion B?". This results are compared in pairs, resulting are in scores and weights. For each pair of criteria required comparing the importance of the two, associating a score as follows (Table 1):

TABLE 1 PAIR OF CRITERIA

Definition	Intensity of importance
Equally important	1
Moderately more important	3
Strongly more important	5
Very strongly more important	7
Extremely more important	9
Intermediate values	2,4,6,8

Numbered intermediate values can be used to define nuances among the five basic formulation. Of course, if it is considered that B is very strongly more important than A, when the opposite is true, so A is assigned the value of 1/7 compared to B. Therefore, since it is assumed that judgments

are consistent with respect to all pairs and all the criteria are "equally important" to themselves, the total number of evaluations will be:

$$\frac{1}{2} \times n \times (n - 1)$$

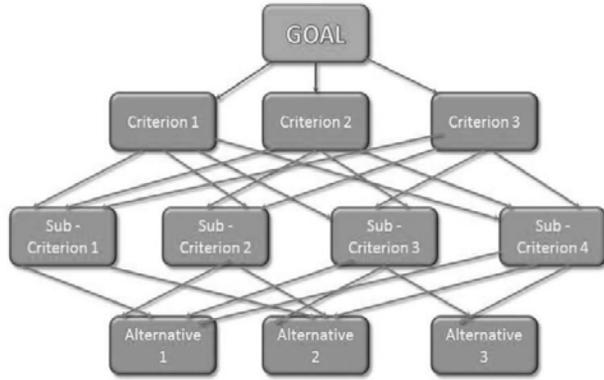


Fig. 2 General hierarchical model in AHP

Application of AHP method can be explained in four steps:

1. Setting a hierarchical model of decision problems in order with goal on the top criteria and subcriteria at lower levels, and alternatives at the bottom of the model (Fig. 1);

2. At each level of hierarchical structure each elements of the structure are compared in pairs, whereby the decision makers express their preferences with the help of appropriate scale which has 5 degrees and 4 sub-degrees of verbally described intensities and the corresponding numerical values for them in the range from 1 to 9 (Table 1);

3. Local priorities (weights) of criteria, sub-criteria and alternatives at same hierarchical structure level are calculated through appropriate mathematical model and afterwards they are synthesized in total priorities of alternatives;

4. Implementation of the sensitivity analysis for final decisions.

The matrix A has special features (all of it's rows are proportional to the first row, and they are all positive and $a_{ij} = 1/a_{ji}$ is true) and because of that only one of it's eigenvalue differs from 0 and is equal to n .

If the matrix A contains inconsistent estimates (in practical examples almost always), weight vector w can be obtained by solving the equation $(A - \lambda_{max} I)w = 0$ with prerequisite that $\sum w_i = 1$, where λ_{max} is the largest eigenvalue in matrix A . Because of matrix A properties $\lambda_{max} \geq n$, the difference $\lambda_{max} - n$ is used in measuring consistency. With consistency index $CI = (\lambda_{max} - n)/(n-1)$ measure of consistency can be calculated:

$$CR = CI/RI$$

$$A = \begin{bmatrix} 1 & a_{12} & \cdots & a_{1j} \\ a_{21} & 1 & \cdots & a_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} \end{bmatrix}$$

Fig. 3 Matrix A

The next step is to determine the set of weights that are most consistent with the estimates of the relative importance of the criteria. This can be done in several ways. In the method developed by Saaty [2], the calculation of the weights is based on a relatively complex mathematical apparatus, using matrix algebra tools. The results are values associated to eigenvector of maximum eigenvalue matrix.

The calculations are quite complex, so it is necessary to use a dedicated program.

But in practice, we provide a simple method of calculation, which gives the same result with two decimal places:

- Calculate the geometric mean of each row of the matrix.
- It calculates the sum of the geometric mean.
- Normalized geometric mean.

IV. RESULTS AND DISCUSSION

The system allows both measurements using dedicated equipment, and with the help of sensors and acquisition board. Measurements can be performed with the help of the mobile laboratory.

Monitoring points were established to evaluate the impact of road traffic on environment and implicitly on people.

By example, in the system, concentration measurements of carbon monoxide can be achieved using MG-811 sensor. Also this sensor is good for measurement CO concentration. This is an onboard signal conditioning circuit for amplifying output signal and an onboard heating circuit for heating the sensor. The MG-811 sensor is basically a cell which gives an output in the range of 100-600mV (400—10000ppm CO₂). The LMC662 is used as the amplifier because of its ultra-high input impedance. According to the datasheet of MG-811, this sensor require an input impedance of 100-1000Gohm. This sensor need one analog input and one digital for connect the sensor to Arduino Mega Microcomputer. The Arduino Mega is a microcontroller board based on the ATmega1280. It has 54 digital input/output pins (of which 14 can be used as PWM outputs), 16 analog inputs, 4 UARTs (hardware serial ports), a 16 MHz crystal oscillator, a USB connection, a power jack, an ICSP header, and a reset button. It contains everything needed to support the microcontroller; simply connect it to a computer with a USB cable or power it with a AC-to-DC adapter or battery to get started.

The block schema is presented in figure 4.

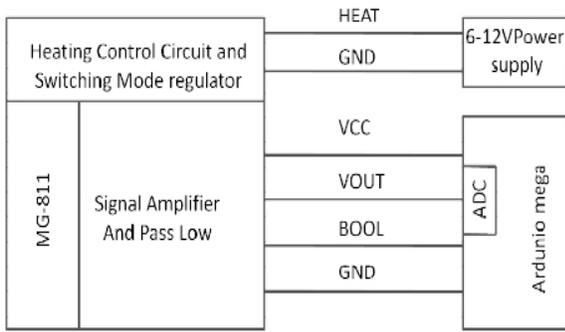


Fig. 4 The block schema of connection, sensor to microcontroller

The data obtained through hardware presented will be analyzed in Labview software. This permitted:

- Easy used Arduino digital input/output, analog input, I2C, and Serial Peripheral;
- Interface from LabVIEW;
- I/O engine sketch to load on Arduino;
- Communication wireless via XBee or Bluetooth;
- Loop rates: USB tethered (200 Hz) and wireless (25 Hz);
- IDE arduino sketch and LABVIEW toolkit VIs help to specification functionality.

The panel of this virtual instrument is shown in figure 5.

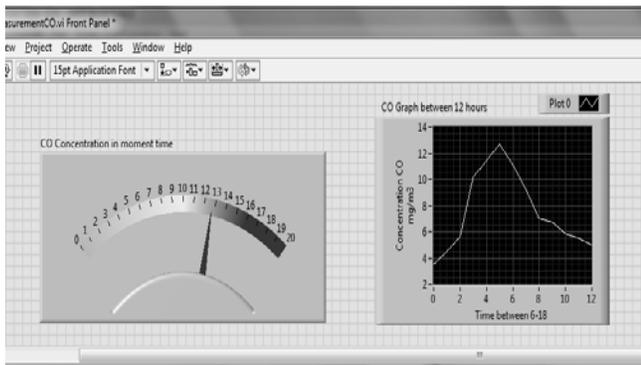


Fig. 5 The panel of CO concentration part of the system of automatic acquisition, processing and analyze data

Furthermore, we will present one of the most important parts of the automatic acquisition, processing and analysis of environment data system, which is the analysis of the AHP multi-criteria method implemented within the system. The virtual instrument created in this system is based on objected-oriented programming. The input data is processed by using the LabVIEW functions. With the help of these functions the data is memorised in a database, in order to be further used in the realization of the objected classes for the generation of the intelligent maps of the Geomedia Professional Software.

The system allows measurements for the diverse polluting chemical factors, like: NO_x, PM10 dust and ozone.

The multi-criteria analysis that we will further present it is based on real measurements of the NO_x, PM10 dust and ozone made in 4 different points from Sibiu: Union square, Alba-Iulia street, DN 1 306 km and the Sub Arini park.

The virtual apparatus projected for the realization of the

multi-criteria analysis works as follows:

- A subjective appreciation of the importance of each point stated above is made depending on the number of vehicles, number of people that are found in that concrete point at a certain time. The comparison matrix of pairs is realized, the weight of each observation point is calculated, the priority vector Lambda, CR and CI are being calculated. The block diagram are presented in figure 6.

- Depending on the measurements made for each of the polluting factors, the virtual apparatus will calculate the weight of each polluting factor from the observation points taken into account, figure 7.

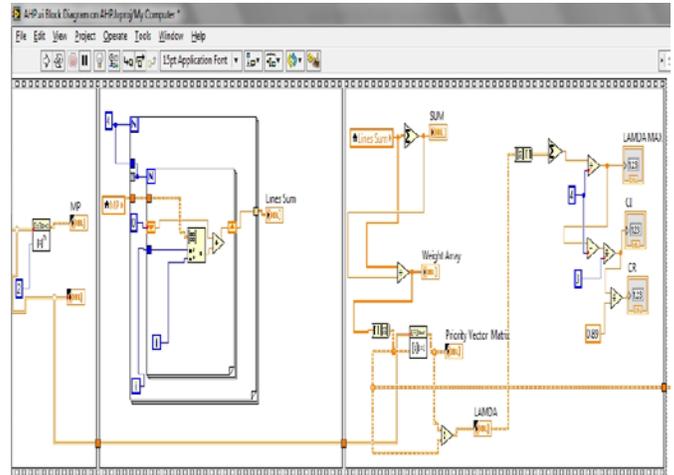


Fig. 6 The diagram of VI for measurement points in AHP analysis

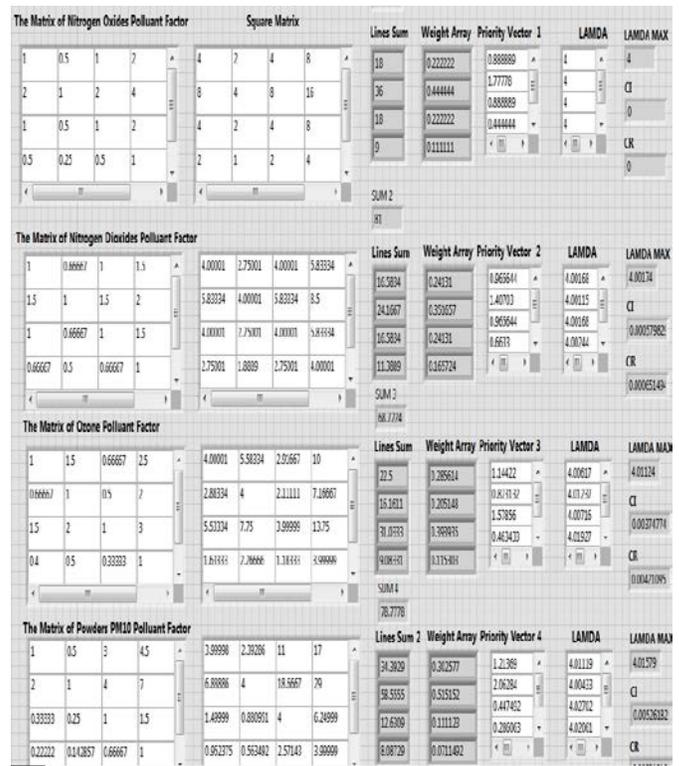


Fig. 7 The panel of VI for all pollutant factors

- Finally, the matrix of polluting factors is obtained. This

will be multiplied with the observation point's weights matrix. The virtual apparatus allows the determination of the most and least polluted observation point, depending on the subjective appreciation made upon them and also, depending on the weight that each polluting factor has in the observation points. The final results are presented in figure 8.

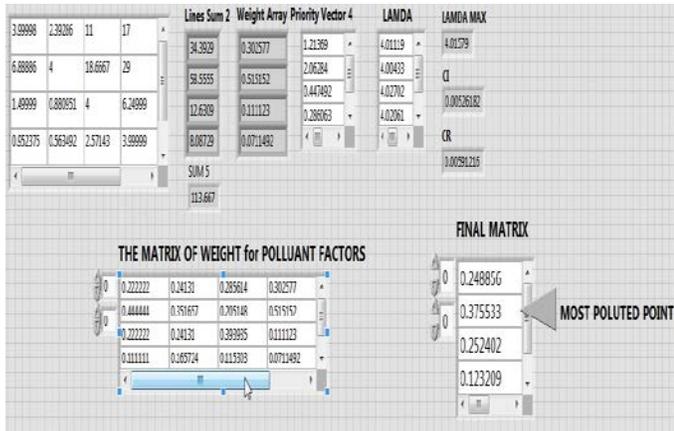


Fig. 8 The matrix of weght for pollutant factors and final matrix with results

For the presented example we can observe that the most polluted point is DN 1 306 km, followed by Alba Iulia street, Unirii Square and, as expected, Sub Arini park.

The presented system allows the automatic memorization of data measured and processed in ACCESS database. This is very important for the GIS Maps generation. The ACCESS table in which the data is saved is in this way projected, as for it to hold the specific object class attributes, which will be presented in the map that will be generated using the Geomedia Professional Software, in figure 9.

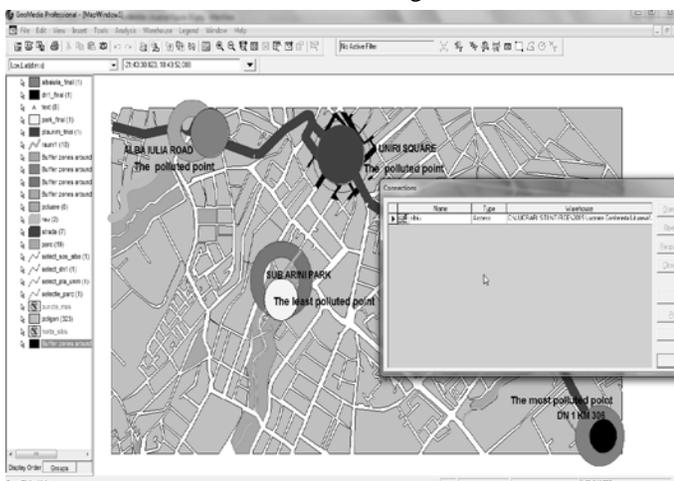


Fig. 9 The map with the observation points and database connection

V. CONCLUSIONS

The automatic system for the taking over, processing and data analysis allows obtaining the object class attributes used

for generating the intelligent maps.

The data are taken over and processed using the objectual technology and the implemented functions in LabVIEW. Plus, the system offers the possibility of the data analysis through sensors, microcontrollers, acquisition boards and specific apparatus, which is highly unused in the current literature. The multi-criteria analysis is made depending on the user's will. The elaborated system is very important because of the fact that it excludes the intervention of the human factor in the acquisition process, taking over and data memorisation.

In the future the system will be extended by using other analysis methods, like for instance the TOPSIS method or other optimization technics based on genetic algorithms. Also, it is very important the automatic generation of the pair comparison matrix, depending on the factor's values

REFERENCES

- [1] T. L. Saaty, "The Analytic Hierarchy Process", McGraw-Hill, New York, 1980.
- [2] T. L. Saaty, "Decision making with the analytic hierarchy process", *Int. J. Services Sciences*, Vol. 1, No. 1, 83-98, 2008,
- [3] B.G. Hermann, C. Kroeze, W. Jawjit, 2007, "Assessing environmental performance by combining life cycle assessment, multi-criteria analysis and environmental performance", *Journal of Cleaner Production* 15 (18), 1787-1796.
- [4] A. Maliki, G. Owens, and D. Bruce, "Combining AHP and TOPSIS Approaches to Support Site Selection for a Lead Pollution Study", *2nd International Conference on Environmental and Agriculture Engineering, IPCBEE vol.37 (2012) © (2012) IACSIT Press, Singapore, 2012.*
- [5] R. Bitter, T.Mohiuddin, M. Navrocki, "LabView Advanced Programming Techniques", Boca Raton: CRC Press LLC, 2001
- [6] D. Beck, H. Brand, "Control System Design Using Labview Object Oriented Programming" Proceedings of ICALEPCS07, Knoxville, Tennessee, USA, 84-86, 2007
- [7] Balogun, I.A., Adegoke J., Carbon Monoxide Concentration Monitoring in Akure. A Comparison between Urban and Rural Environment, *Journal of Environmental Protection*, 5, 266-273, 2014;
- [8] EPA, (2014), Clean Power Plan, Available: <http://www.c2es.org/federal/executive/epa/q-a-regulation-greenhouse-gases-existing-power>;
- [9] Vibhor S., Shivani S., Rajesh B., Traffic related CO pollution and occupational exposure in Chandigarh, India, *International Journal Of Environmental Sciences* 5(1), 170-180, 2014

Port operation – increase of automated systems, decline of workforce jobs?

Aureo E. P. Figueiredo, Ricardo de D. Carvalho, Sérgio Hoeflich, Letícia Figueiredo, Sergio L. Pereira, Eduardo M. Dias

Abstract— As ports are worldwide centers of intermodal changing in transportation of goods, since last years of the nineteenth century are increasingly equipped with lifting and cargo transportation equipment. By this way, both the massive mechanization of the activities and the evolution of automation technologies opened the window of opportunity to change processes and use of systems with automatic controls, as long as the number of dockworkers decreases. In several industrial plants, the effects of these changes are quite visible, as long as result in increased production and lower risk exposure. However, from the point-of-view of jobs opportunities, the inverse is not strictly proportional, in comparison with significant people reduction. The use of robots in industrial manufacturing led to the emergence of so-called *lights out* processes or *dark factory*. Considering the new trend in progressively mechanized/automated cargo terminals, the displacement of port workers for other activities is very real and intense. Regarding some peculiarities and characteristics of multipurpose terminals handling in ports, will this trend dominate for the foreseeing future? This article discusses this question in three parts: automation, port automation and the effects of mechanization, automation and possible consequences for jobs and work in ports.

Keywords— Automation, Port Automation, jobs and workforce.

I. INTRODUCTION

THE cargo movement in ports undergone a worldwide transformation. The infrastructure of port waterfront, berths, facilities have evolved considerably, making it

Aureo E. P. Figueiredo is with GAESI - Grupo de Automação Elétrica em Sistemas Industriais, a research group of the Electrical Energy and Automation Department, Polytechnic School of University of São Paulo, Av. Prof. Luciano Gualberto, trav. 3, n. 158, São Paulo/SP, Brazil, CEP 05508-970 (e-mail: aureo.figueiredo@gaesi.eng.br).

Ricardo de D. Carvalho is professor at the School of Engineering of Santa Cecilia University - Unisantia/Santos. (ricardo.carvalho@ogmo-santos.com.br)

Sérgio Hoeflich is with GAESI (shoeflich@ig.com.br)

Letícia Figueiredo is a MSc student at Polytechnic School of University of São Paulo (leticia.figueiredo86@gmail.com)

Sergio L. Pereira is professor at the Polytechnic School of University of São Paulo and a researcher of GAESI (sergioluiuze@uol.com.br)

Eduardo M. Dias is full professor at the Polytechnic School of University of São Paulo and coordinator at GAESI (emdias@pea.usp.br).

capable of supporting efforts from load requests of thousands tons arising from equipment appliances mounted.

The mechanization and the introduction of container led to the reformulation of shipping transport and the whole operation in ports. The automation of port systems brings continuous advances. Cargo handling reach increase amount meanwhile, for workers, it is observed a severe reduction of jobs. This paper addresses this issue in three stages: automation, port automation port and discussion of the effects on dock work.

According to historical records dating back to 270 BC, Vitruvius studied possibilities of a clock powered by water, apparatus referred to as rudiments of automation.

In productive activities, it is considered replace work and human efforts by devices and equipment that have better overall result, concept called mechanization.

In an applied way, we have the progressive use of tools and systems based on physical principles related to: work, momentum, power, hydraulic and pneumatic pressure, wind, gravitational, thermal, electric, solar power, etc...

The development of mechanisms and equipment and its incorporation into the organized production dates back to the period of so-called Industrial Revolution, with the classic example of the looms with drives the steam produced in boilers and mechanical sets of motion transmission as cylinders, piston, piston rods, shafts, pulleys, rotors, cams, distribution gears, etc.

According to Mamede Jr [1], "from the industrial revolution, muscle strength gave way to machines," by mechanized production.

The systematization of productive activities allowed the occurrence of the second industrial revolution, the implementation of the assembly line, in the early years of 20th century, with the large-scale production of the products.

Therefore is widely accepted that the process of mechanization is a preliminary step in gradual trend on enroll in to automation systems.

Automation sets the basis for the third industrial revolution, the revolution resulting from the digital age. Thus, in a reverse cycle, a large scale production provides a possibility for manufacturing using high technology even for few quantities with compatible costs, starting with sophisticated sets as jet engines.

II. INDUSTRIAL AUTOMATION

Some authors indicate Automation is a neologism from automation, and refers to the use of technology to facilitate the work of human or extend their physical and mental capacity.

Control and automatic control refers to the use device (controller) that without the help of human action make a system behave the way you want.

Among the areas of knowledge involved in automation, we have the inter sciences: mechanical, electrical, electronics, telecommunications and information systems.

This consideration is quite important, as they relate Zugge, Pereira, Dias [2], "the problem is frequent considerable that a lot of information does not create any knowledge". In addition, recommend for the improvement of systems the integration between IT (Information Technology) and AT (Automation Technology), "Nowadays some companies are investing a lot of money and effort to integrate IT and AT. However, some companies have not been achieving the success they expected "which indicates the complexity of this improvement".

III. HIERARCHY LEVELS OF INDUSTRIAL AUTOMATION

Industrial automation can take various roles in industries, since the automation of an operation to a large set of sequential activities, with the participation of so-called industrial robots.

Facing to the diversity of production processes, there are different types of classification called Automation Level. The table below shows the traditional model Purdue Reference Model, levels basis for the standard ISA 95.

Table 1 - Purdue Reference Model PRM

Levels	Description
5	Business Systems, strategic planning and corporative management
4	Plant level, production and programming, ERP, MRP and MES
3	Operation Unit Level
2	Machine/Process Automation Level
1	Controller Level
0	Sensor/Actuator level

Lydon, B [3] points to a technology evolution that could simplify that level hierarchy, as "more controllers are supporting multiple Ethernet ports to interact directly with industrial and business network that exists throughout industrial plants".

In other way, Rother M Harris R. C , [4] describe a hierarchy that consider systems subdivision with different levels of automation in its subsystems or processes, also said semiautomatic, where there is partial involvement of manual action.

Table 2 - Adapted from Rother M. Harris R., LEAN directions

L E V E L S		Machine Load	Machine Load	Machine Discharge	Transfer Parts
	1	Manual	Manual	Manual	Manual
	2	Manual	Auto	Auto	Manual
	3	Manual	Auto	Auto	Manual
	Main Division				
	4	Auto	Auto	Auto	Manual
5	Auto	Auto	Auto	Auto	

And also explains "level three offers the most efficient and flexible combination of the movements of the operator and materials."

IV. PROCESS AUTOMATION – INDUSTRIAL PLANTS

The market competition makes the automation of processes in the industrial area a very important condition, with the increasing deployment in the assembly lines of automated procedures, using robots and mechatronic devices.

By the automation of production, the industrial sector has advanced considerably with the introduction of robots in assembly lines.

In a compact physical space of the production line, robots replace humans with speed and efficiency, especially in repetitive and dangerous tasks with greater potential risk of accidents.

The nature for the related process has distinct impacts behold production conditions usually defined in industrial plants.

For production planning, the limitation of time and space, with constant repetition of movements, paths and routes are likely to be supplied and standardization within the degrees of freedom of movement of robots.

An important reference is called standardized time, or simply *takt time*, that Rother M, Harris R. [5] at LEAN methodology describe as "the rate at which the finished product needs to be completed in order to meet customer demand," so, described mathematically, *takt time* is:

Available time for production / required units of production

Nowadays, the *digital hortator* beats by software and hardware the rhythm of industrial operations.

V. PROCESS AUTOMATION – INDUSTRIAL PLANTS, DARK FACTORIES

Many industrial manufacturing units are named manufacturer, referring to the Latin word for to handmade, or handcraft, i.e. to make something by manual skills. Industrial sector uses strongly automation on the assembly lines considering the usage of robotics, mechatronic devices, pneumatic systems and other types of automated systems.

Conditions defined for industrial plants as production and

process characteristics favored this transformation.

The growth of this production system induces, at its limit, to a highly robotized production environment, where the man participation is minimal, practically restricted to the supervision and control of the continuity of processes.

This automation, robotics where the drive eliminates the lighting, result "manufacturing lights out" processes, and industrial plants also known as "dark factory". The need for lighting is the human controller of the process, at least for repairing, adjusting jam although mainly focusing on system supervision.

This situation was not yet effective for several reasons, including the complexity of the robots and their high costs of implementation and operation, requiring expensive specialists.

Still, some areas of production such as plastic injection, laser cutting and printing have made significant progress in automation.

The increased use of robots in factories will be possible with the advancement of technology, allowing less expensive costs of acquisition and set up robots, in order to simplify the operations in great scale, and even more productive.

Currently, industrial production processes still consider as high valuable the participation of man, in new tasks in which the abilities and skills are useful.

VI. PORT AUTOMATION - INTRODUCTION

The application of automation in port processes in the movement of goods led to structural changes.

As well known, the transportation of goods from the coast to the vessels along centuries was performed by smaller boats - with the climatic difficulties of sea currents and waves.

This situation stimulated, where possible, the construction of precarious wooden structures to where the draft allowed positioning the vessel, for a dry access to the ships.

The building of safe pier docks with solid structures in areas protected from the weather (and also from pirates), brought the ship moored closer to waterfront, accessible platform for animals and land transport vehicles, and implementation of deposits near the berth, adjusting production to ship traffic.

The stable bases permitted to developed specific equipment for lifting and transport of loads that, according P. Alfredini, E. Arasaki .[6] should consider the continuity of operations without times waist in intermediaries of equipment operations, by the use of temporary storage areas with multifunctional devices.

VII. AUTOMATED VESSELS

The continuous advancement in technology brings every day new automated ways in the movement of goods. Therefore, companies started to design cargo carriers without onboard crew, considering remote operation.

This condition has resistance in several countries, facing to the current lack of security guarantees, mainly environmental.

On the other hand, presents itself as a perspective for the near future, operating in the marine environment like the drones aerial vehicles VANTs (unmanned) or VARP

(remotely piloted).

VIII. AUTOMATED TERMINAL

In containers terminals, the fundamental rules of logistics guide the design of load plans by the ship planners, considering the sequence of ports at which the ship will operate, according that "the first a load is boarded the latter will be removed".

The relative position of the containers in the ship hold, (inside), or on the deck (outer), is referenced spatial orientation relative to the ship (bow, stern, portside, starboard), according to the position set on the loading plane array. Bay (longitudinal) and row (transverse) consider a horizontal plane, and the vertical tier position.

Fig. 1 presents the bay, row and tier positioning.

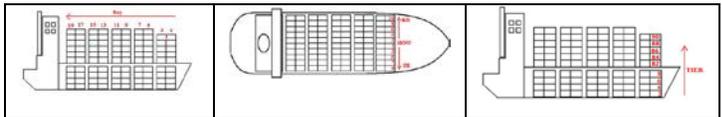


Fig. 1 - Bay, row and tier positioning. Containers loading matrix. Image adapted from Oceanica UFRJ.

Regarding its capacity containers, vessels are indicated as shown at table 3.

Table 3 - Adapted from oceanica UFRJ

Number of TEUs	Vessels Types
< 1000	Small feeder
1001 - 2800	Feeder
2801 - 5100	Panamax
5101 - 10000	Post Panamax
10001- 14500	New Panamax
> 14500	ULVC Ultra Large Container Vessel

Nowadays one of largest ship in operation at the time, CSCL Globe, with a capacity of 19,000 TEU, moored in Rotterdam in January 2015. It also announces the construction of a new ship for more than 20 thousand TEU.

For the East-West route, the main in terms of cargo volume, these super ships are supposed to operate mainly in the future channel of Nicaragua with planned capacity for vessels up to 25,000 TEUs or 400 thousand tons. The Panama Canal, even with the current expansion, had its capacity increased up to 15,000 TEUs or 150,000 ton (New Panamax).

IX. AUTOMATED OPERATION IN CONTAINER TERMINAL

When ship is moored at the pier and its position is validated, this condition also validates up the spatial position of each container indicated in the matrix structure of the cargo plan, allowing the start of operations.

From the ship plan, based load plan, along with the movement of the lifting equipment, an automated operation the "carousel" of trucks driven by programming can start the

journey with stroke and established operation, integrated with other equipment such as reach stackers, RTG (rubber tired gantry crane) RMG (rail mounted gantry crane) transporters and quarry gantry cranes.

Port operations Multi cargo ports:

Producer to → Regulator yard → Port Terminal → ship
Generic boarding algorithm

Port operations could be realized in many levels of mechanization and automation. Equipment considered for each material and mode of transport, road, railway, waterway and pipeline. In smaller scale, the use of lifts and airport connections systems.

X. OPERATION WITH CONTAINERS – GENERIC BOARDING ALGORITHM

Producer to	Regular Yard	Port Terminal	Ship
Truck Arrival Notification	Standby Scheduling Authorization Continues Trip	Access Identification Positioning Control Equipment Discharge (forklift/reach stacker /rtg) Loading Truck Transport Positioning Quarry Crane/LTM	Receipt Positioning Bay/Row/Tier Lashing Liberation Ship Deliver

This is a type of mechanized operation with good prospects for automation in ports and terminals.

Since receiving the container in the terminal entrance, access control, positioning in the courtyard, integration into the cargo plan, the mobilization to the sideline, loading on board in ship hold or deck.

On arrival at the port of destination, repeat the sequence in reverse order, according to the containers schedule.

XI. OPERATION OF DRY BULK – GENERIC BOARDING ALGORITHM

Producer to	Regular Yard	Port Terminal	Ship
Truck Arrival Notification	Standby Scheduling Authorization Continues Trip	Access Identification Positioning Control Equipment Discharge (Hopper/Elevator/conveyor/Tripper) Load (Carrying Shovel) Hopper/Elevator/conveyor Ship loader	Cargo Hold Receipt Liberation Ship Deliver

The solids unpackaged, bulk are generally derived from products:

- Vegetable - sugar, cereals (maize, soybeans, wheat, and others) pellets. These products are dependent of protection against moisture and must be packaged in appropriate warehouses.
- Minerals: sulfur, phosphate rock, bauxite, salt, coal and raw materials for fertilizers among others. According to the particle size and other physical characteristics, the products could stay in a open yard.

The packaging of bulk solids requires protection systems and environmental control, and workers. According to the particle size and other physical characteristics, plant products are flammable and require protection systems and firefighting.

Ships for bulk solids transport must be equipped with the appropriate holds to receive loads, busy with cargo equipment or continuous discharge or batch.

Table 4 describes capacity for bulk carriers.

Table 4 - Capacity for bulk carriers. Adapted from http://www.worldtraderef.com/WTR_site/vessel_classification.asp

Capacity dwt	Vessels Types
10.000 to 30.000	Handy Sized
30.000 to 50.000	Handy max
50.000 to 80.000	Panamax
80.000 to 200.000	Capesize
>200.000	Very large ore carriers
400 mil ton	Valemax

The bulk solids for export is received at the terminal from trucks, waterways barges or rail cars with discharge hoppers.

For the unloading of trucks is often the use of the truck lift system, and gravity unloading the posterior position of the cart.

Railway wagons have bottom opening and systems to discharge directly into the hoppers.

The hoppers are willing transported on belt conveyor for yard or warehouse conforming trapezoidal piles from mobile discharger from a structure tripper.

Another form of storage are warehouses of grains, generally circular and high dimensions that generates concentrated loads, which require special soil foundations.

At the boarding operation, bulk is loaded with belt conveyor that carry the products directing to the holds of ships equipment called ship loaders. When are used bags the ship loaders are provided with a dispensing reel.

In reverse operation, the ship unloader, the discharger removes the vessel's hold material grabs or suction, according to the product characteristics. Hence, after discharge into a hopper, it is transported to the warehouse or directly to trucks and wagons.

XII. LIQUID BULK – GENERIC BOARDING ALGORITHM

Producer to	Regular Yard	Port terminal	Ship
Ducts Notification Pumping	Receipt Stock Scheduling Connection Ducts Notification Pumping	Receipt Stock Scheduling Connection Ducts Notification Pumping	Cargo Hold Receipt Liberation Ship Deliver

Traditionally, liquid bulk is one of the most automated forms of goods handled in ports. The predominant cargo are fuels of mineral or vegetable origin, vegetable juices, liquefied gases, chemicals.

They always are careful operations that require strict compliance with safety standards to avoid disastrous occurrences.

The placing product in tanks requires holding basins for any leaks, level control devices, pressure, and temperature. In the discharge and suction operations, the safety valves have special care, as well as grounding systems and prevention systems, monitoring and automated firefighting.

It is known that even small electrostatic charges may cause sparks and harmful consequences in environments with potentially flammable materials and explosives.

The liquid bulk from tanks is pumped into the vessels hold; all with the preventive measures the terminal itself or port PAM - Mutual Aid Plan - when necessary Publication Principles

XIII. GENERAL CARGO - GENERIC BOARDING ALGORITHM

Producer to	Regular Yard	Port Terminal	Ship
Truck Arrival Notification	Standby Scheduling Authoriza- tion Continues Trip	Access Identification Control Equipment Discharge (Forklift/Reach Stacker/Rtg) Loading Truck Transport Positioning Quarry crane	Receipt Positioning Bay/Row/Tier Lashing Liberation Ship Deliver

Products not classified as containers or bulk are referred as general cargo.

Several products are moved packed in bags of various sizes. More recently, in large bags. Also drums, coils, profiles, ingots, bales, etc .

Transport could be combined on wooden pallets or tied pieces or slings

Still under the general cargo designation are included big parts of project cargo or special sets or large unit equipment, volumes or masses, such as turbines, electric transformers, rotors, locomotives, cranes, propellers of wind generators, quarry cranes, RTG cranes, wagons, locomotives, rails, ships, with brackets and specific lugs for connection of lifting and conveyance. If necessary, are used specific vessels for special transport.

For this study highlights the cargo of wind propellers, the difficulties lifting, positioning and locking of packaging. There is no automation in this operation.

XIV. PROSPECTS FOR AUTOMATION – IN ADDITION TO EFFICIENCY, SAFETY AUTOMATION IN TO THE PORTS

About security in domestic and foreign concept of a port or terminal, highlights the IHMA- International Harbour Masters Association [8] is:

“Most, if not all, navigable rivers, channels, ports, harbors and berths are subject to danger from, for example, tides, currents, swells, banks, bars or revetments, traffic density and changes in depths.

And continues: “Such dangers are frequently reduced by lights, buoys, signals, warnings and other aids to navigation and can normally be met and overcome by proper navigation and the handling of a vessel in accordance with good seamanship. And concludes: “The reputation of a Port is largely dependent on its safety record and efficiency. Any damage to a port’s safety record may affect on its reputation and by extension, its trade” Many ports have their own VTS – Vessel Traffic System, for controlling arrive, port activities and depart of ships.

About safety, the commotion caused by accidents and personal injuries must be considered as the port environment

influences directly to the worker safety and integrity, his family, the work team, and even relationship social circle, feels the impact of any accident victimized by the worker.

Accidents tend to cause breakdown of the teams motivation contaminating the work environment, underscoring due to their frequency and severity.

However, how society acknowledges and judges accidents?

Thus, the moral and social losses are exacerbated by the ratio of fatalities and perception, as presented in IMO-FSA [9], which demonstrates the public aversion to accidents when he says:

"Society in general has a strong aversion to accidents with multiple victims. There is a clear perception that a single accident victims in 1000 is worse than 1000 accidents in which dies (kills) a single person."

If, on the one hand there is a spectacle more severe injuries, on the other notes to popularization of smaller, affecting the balance of his analysis and correction. Mechanized equipment and automated systems must have operation conditions that preserve people that operate them.

XV. PORTS, AUTOMATION AND PRODUCTION WORKFORCE

A. Socioeconomic aspects and conflicts at work.

The implementation mechanisms with increasing production capacity have correspondence with the reduction of jobs in these activities. This situation led to armed conflict as referred to historically playfulness, social movement in the second decade of the nineteenth century.

Led by Nelson Ludd, aggregated workers who mobilized against replacing workers with machines, invaded industries and breaking equipment. Afterwards, unions appeared to collectively fighting over workers' rights.

B. Automation and work

On systems with self of automation, should be avoided human presence restricted to authorized circulation areas, so should not remain people to avoid accidents, even with sound from handling equipment when moving.

It is what is observed at the terminals where automation is implemented progressively reducing up the human presence in the courtyards. These act in the control and supervision of the process steps. Centralized automation system provides a possibility for the operator to monitor and control field area by checking high-resolution cameras. An actual example is the Euromax terminal in Rotterdam.

These terminals due to their level of automation seem like ghost terminals, where is not allowed people stay in the cargo area.

Its main feature is the reduction and even elimination of direct human participation in the activity, with the man in the monitoring and supervision. It follows from the cost savings resulting from energy saving, and materials, improving the accuracy and quality of operations.

The automation of operations and processes frees man's work in repetitive activities and unfavorable conditions in

which there is discomfort and potential risk of accidents and illnesses. The reduction of costs related to the work reflects the lower application hours and related charges.

So with less human involvement, expected better quality and productivity, with fewer accidents due to fatigue and tiredness. Additionally fewer complaints and wage demands.

However, even in the most automated processes in dark factories, have not yet reached projected results, the basic finding of the operational difficulties of adjustment processes, which confirms the finding that the most important in the production process are people prepared to realize it and driving it.

Hence the contribution that experience and knowledge are fundamental to success in the processes. They should be the subject of care in their preservation and enhancement for transmission to reviewers.

C. Port jobs and workforce

Process automation - terminals and non-specialized port facilities

The goods handling environment in the ports of various loads differs significantly, almost in opposition of the processes in industrial plants, since in them notes the diversity of goods, work areas, actions to develop and mainly of agents who perform, workers.

For the workers, every period change the places where carry out the activities, ships to be operated, the corresponding goods and procedures, co-workers, the hot / cold climatic conditions, the periods of day / night, natural light or artificial, or even the night pitch of the paths between rows of containers.

In cases where partners and staff members are permanent, it is natural to develop collective skills, interaction for knowledge, in synergy characteristic of team game that allows influence for improvement.

In these teams, alternating hierarchy at every turn, knowledge also requires frequent training to perform multiple tasks, lead and organize.

The intensive use of human labor in working conditions under hard working conditions, characterized long period until start the utilization of large equipment mechanization that caused the deployment of people.

At the beginning of the use of clamshells as implements of cranes for bulk handling, tells us T.C. Zotto [7] about the gunshots from angry workers in the first grabs, similar to actions such as the attacks and destruction of equipment of the industrialization period.

From the port worker point of view, even arduous working conditions are important to support subsistence of themselves and their families. Hence the historical reaction to processes that reduce jobs.

However, for a long time, the man remained the major force element in the movement of goods on board ships and ashore.

The use of lifting equipment and cargo handling evolved the capacity of quickly load ships to increase the number of trips and dispose of goods to destination markets. This evokes the observation of K. Marx [10] in the mid-nineteenth century:

"You can ask if all the mechanical inventions made so far lightened the daily work of any human being", since the true intention of the capital was to reduce the price of the goods.

In contemporary terms, T. Picketty [11] manifests one of the expressions of this bias automation in his work "The capital in the XXI century", which argues that (r) rate of return on capital will fall slower than (g) rhythm economic growth.

Hence, if it is sufficiently easy to replace workers with machines, if the elasticity of substitution of capital for labor is greater than one, widen the gap between r and g . thus concludes an immediate consequence of redistribution of income, workers in return for capital.

D. Changes in dock work - Containers, production and jobs

Port work at early years of the twentieth century was still essentially manual. Huge queues of workers carrying bags in the back were the typical views of labor between warehouses and ships.

At the end of the twentieth century and the beginning of the current, was the multiplication of cargo movement, fostered by the dynamics of the economic situation and facilitated by the automation of processes, with the "revolution" world due to the intensive use of containers, with loads of all kinds put in charge safes.

This conceptual change leads to what F. Bruno et al. [12] refer to "the simple thing the container did was sharply lower the cost of shipping goods from one place to another. However, the container revolution also changed the mechanics of shipping: the logistics, the speed and capital's structure".

Add that the use of containers provided for the conditions, and the mechanization of the operations, the automation of the terminals.

"that containerization is a monument to most powerful law in economics, that of unanticipated consequences".

And about jobs reduction, continues Fabiano [12]: "labor leaders feared the container, but even they were not prepared for the speed with which destroyed "water-front-job", against the prediction of 30% reduction, in reality 75% has vanished by 1976".

Port activities implementation are traditionally held by temporary contracted workers, strongly aggregated into their own unions, with great tradition of defending their categories.

How then place the reduction of temporary dockworkers? The Brazilian law of port modernization has established the creation of OGMO - Organization for Labor Management, by replacing the workers of the Port Authority. In addition, policies for training and redeployment of workers.

Modern equipment has operational characteristics that allow the operator to go through a preliminary training simulator to know the movements and possibilities. Then, when it is demonstrated the ability (skills) enough, supervised training takes place in own equipment under limited conditions to acquire the resourcefulness needed to take the solo operation.

Equipment manufacturers developed simulators systems with workplaces similar to the operating equipment, sensors, controllers and speed adjustment devices, range, and other

control measures of the operation.

Thus, the increasing mechanization and automation in port and multipurpose cargo terminals are incipient, and progressively increased by the incorporation of technologies directed to the speed and intensity of movement, creating complex situations that require constant attention.

XVI. CONCLUSION

About automation influence in production activities, McMillan [13] observes that: "People make errors. Automation has been very successful at reducing these errors although it may just be relocating human errors to another level."

In addition, continues, "A person is still needed for performing cognitive-based tasks, as another system check and provide needed flexibility for unexpected events".

When people experimented is been pushed away from the workforce, more specialists systems are required, to support decisions. The results may be not the same.

Thus, for multimodal terminals and ports, the present situation is certainly still far away from the so-called *ghost terminal*, where intense automation and consequent reduction - *invisibility* - the absence of human work would be the compared to *dark factories* of industrial activities.

If, for advanced cultures is understood that automation displaces people from dangerous function, painful and exhaustive, more and more the training people to work in automated systems is required for repositioning of workers.

Ports applications of automated systems must preserve for a long time the workers contribution, based in skills and experience, ability for prevent and problems solutions.

REFERENCES

- [1] J.Mamede Jr, "Instalações Elétricas Industriais", Capítulo XIV Automação, Ed. LTC, 2007, p 633.
- [2] C.T.Y. Zugge, S.L. Pereira, E.M. Dias, Integration of Information Technology and Automation: Facilitators and Barriers". WSEAS transactions on systems and controls, issue 5, v 5, 2010 p 372.
- [3] Lydon, B. "Simplifying Automation System Hierarquies" 2012 Editor Automation.com access march20,2015.
- [4] Rother M. Harris R., "Creating a continuous flow" Publ. LEAN Enterprise Institute, 2001 p13.
- [5] R.Harris, "O nível de automação ideal" acess march 2015 http://www.lean.org.br/comunidade/artigos/pdf/artigo_43.pdf.
- [6] P. Alfredini,;E. Arasaki, Engenharia Portuária, p 770, Ed. Blucher, Brazil 2014
- [7] T.C. Zotto, "O trabalho de Estiva - Modernização e Tradição, os desafios da tecnologia e da gestão no cais, SP LTr, 2002.
- [8] Harbour Master Nautical Safety and Port Environment. Available: <http://www.harbourmaster.org/hm-port-safety.php>. Access: March 2015
- [9] IMO International Maritime Organization - Amendments to the guidelines for formal Safety Assessment (FSA) for use in the IMO Rule-Making Process (MSC/Circ.1023 - MEPC/Circ.392) FN Diagrams oct16 2006.
- [10] K.Marx The Capital Livro 1 n 1, Ed. 70 Lisboa, 1979 p. 69
- [11] T. Picketty The Capital in the XXI Century,
- [12] B. Fabiano, F.Curro, A.P.Reverber, R.Pastorino."Port safety and the container revolution, A statistical study on human factor and occupational accidents over the long period" - Chemical and processes

engineering department "G.B. Bonino" University of Genoa, Italy.
Safety Science 48 (2010).

- [13] McMillan, G.K. "Process Industrial Instruments and Controls Handbook", McGraw Hill 5th ed. 1999, pg 8.18

Implementation of Track and Trace System for Medication in the Largest Hospital Complex in Brazil

Elcio B. da Silva, Maria L. R. P. Dias, Eduardo M. Dias, Sergio L. Pereira

Abstract—This paper aims at presenting a pilot project for the implementation of a system of traceability of medicines in the largest Public Hospital complex in Brazil, Hospital das Clínicas, so that it is able to comply with the regulations of the Brazilian National Medicines Control System. Firstly, it introduces the background of the track and trace regulation in Brazil and its impact on Hospital das Clínicas. Secondly, it discusses track and trace implementation processes in five other countries, comparing their experiences. Then it presents and comments on the specific features and challenges of the Brazilian law, before proposing a pilot project for Hospital das Clínicas. This pilot project underlines the importance of an effective model for the implementation of track and trace mechanisms in the whole medicines supply chain.

Keywords—Traceability of medicines, SNCM, pilot project, RDC54/2013, public hospital.

I. INTRODUCTION

Increasing efficiency in the health care system is one of the main responsibilities of any government. In this process the correct use of medication is a key factor. Today in Brazil there are three predominant irregularities to be faced in order to contribute to improvements in the use of medications. They are the following:

- a. Use of medicines without certified origin, due to falsifications, contraband, and product robbery during transportation;
- b. Sales of medications without therapeutic equivalence, fostered by bonuses given to pharmacies, which are paid by the downstream links of the pharmaceutical chain;
- c. Elimination of unsafe practices of transportation and storage of medicines, which can endanger their quality.

The Brazilian government trying to improve patient's security established medicines traceability across the

pharmaceutical chain with the National Medicines Control System (SNCM), by passing Law 11.903 on January 14, 2009 [1]. Law 11.903 was systematized by the Brazilian National Health Surveillance Agency (ANVISA), through a Board Resolution, Resolução da Diretoria Colegiada (RDC) n.º 54 in 2013 [2], detailed by norms contained in Instrução Normativa (IN) n.º 6 in 2014 [3], and clarified by a technical note, Nota Técnica (NT) n.º 1, in 2015 [4].

The SNCM aims at preventing simultaneously the three main irregularities in medication use. With this aim, ANVISA conceived a unique logic model to regulate the operation dynamics of medicine distribution [5][6].

Hospital das Clínicas of the Medicine School of the University of São Paulo (HC/FMUSP) is the largest public hospital complex in Brazil, with a capacity of 2,200 beds. The HC, as an important agent in the Brazilian health system, must have its operations comply with the demands made by Law n.º 11.903/2009 within the deadlines specified by RDC n.º 54/2013.

Taking into account the pioneer aspect of the logic model proposed by ANVISA, the lack of empirical proof of the effectiveness of available technological solutions and the extent of operations of HC, implementation of SNCM becomes a highly complex activity, which must be carried out without adding costs for the patients or the institution.

As there is no possible confirmation of a mapping of risks, their probabilities, effectiveness of risk-mitigation mechanisms, as well as strategies of reaction to them, this paper proposes, theoretically, a scope of execution of a pilot project for implementation of SNCM in HC.

The execution of a pilot project in HC, supported by a holistic view and scientific rigor, is a valuable tool to accelerate the process of implementation of SNCM in this institution. The pilot project for HC has moreover the potential of collaborating to answer questions about the implementation of SNCM in the public sector, which is one of the main factors of doubt as to the feasibility of compliance with the regulations in the specified deadlines [7].

The scope of the pilot project proposed in this paper was built based on analyses of:

- a. Five implementation experiences of systems with similar purposes (Turkey, Argentina, China, Europe, and the United States), which are pioneer

E. B. da Silva is with GAESI - Grupo de Automação Elétrica em Sistemas Industriais, a research group of the Electrical Energy and Automation Department, Escola Politécnica, Universidade de São Paulo, Av. Prof. Luciano Gualberto, trav. 3, n. 158, São Paulo/SP, Brazil, CEP 05508-970 (e-mail elcio@integradora.com.br).

M. L. R. P. Dias is with GAESI (e-mail: lidiadias@pea.usp.br).

E. M. Dias is full professor of the Escola Politécnica of the Universidade de São Paulo and coordinator of GAESI (emdias@pea.usp.br).

S. L. Pereira is professor of the Escola Politécnica of the Universidade de São Paulo (sergioluizpe@uol.com.br).

projects and represent significant improvements in global health care;

- b. Requirements defined by ANVISA;
- c. Present characteristics of HC in terms of business, operational and technology processes.

The structure adopted in this paper consists of four topics which, in the order now presented, provide the following points:

- a. A comparative analysis of the models adopted globally;
- b. The presentation of the Brazilian model and a discussion of what this Brazilian model implies for the health system as a whole, for HC, and for the patients;
- c. The presentation of HC scenario and the proposal of scope for the pilot project;
- d. Conclusions and proposals for future studies.

II. GLOBAL EXPERIENCES IN THE IMPLEMENTATION OF MEDICINES TRACEABILITY

A. Medicines Traceability in Turkey.

The Turkish medicines track and trace system is called ITS (abbreviation of *İlaç Takip Sistemi* which in Turkish means "Medicines Traceability System"). Bearing in mind that the ITS has been operating since 2010, it can be considered a pioneer implementation. Medicines traceability in Turkey covers all medication commercialized in the country which receive a single mark at item level, i.e., on the medication's secondary package [6][8]. The marking on the secondary package is made by a datamatrix two-dimensional code.

Traceability in the pharmaceutical chain in Turkey comprises manufacturers, distributors, hospitals, pharmacies, assistance organizations that need reimbursement with medication expenses, and the country's health surveillance agency, "Turkish Medicines and Medical Devices Agency", which plays a crucial role in the process, as illustrated on fig. 1.

In the Turkish model medicines traceability control consists of two layers:

- a. Corporate layer: In this layer the organizations which participate in the manufacturing and distribution of medicines have a system architecture dedicated to control: the receipt of valid serial numbers (in the case of manufacturers), the medicines movements within the organization to identify medicines with serial numbers, the permissions from the government layer to transfer the possession of medicines to another organization and, lastly, the report of internal movements considered critical by the surveillance agency.

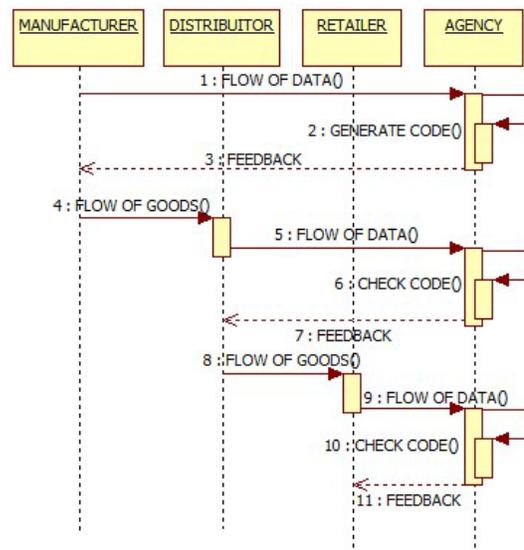


Fig. 1: Physical and logical flow of serialized medicines distribution in Turkey.

Government layer: The government layer is responsible for controlling serial numbers liberation for medicine identification, for validation of transfers of possession among organizations that belong to the chain, for registration of movements, and for storage of track and trace data of medicines commercialized in the country.

B. Medicines Traceability in China.

A "China Food and Drug Administration" (CFDA), the Chinese health surveillance agency, determined the implementation of medicines traceability in China in waves. The first wave, which started operation in 2013, covered serialization of the pharmaceutical chain of medicines listed on the "National Essential Drug Lists" (NEDL) of 2009. After the first wave there followed a series of partial implementations, aiming at total completion of track and trace implementation in December 2015, with serialization and control by CFDA of all the products commercialized in the country [9].

In terms of control structure the Chinese model is similar to the Turkish model, in so far as both consist of two layers. The significant differences dwell basically:

- a. In the way in which manufacturers obtain permission to serialize medicines in the CFDA,
- b. In the set of movements that are controlled by CFDA in each link of the chain, and
- c. In the strategy of implementation adopted.

Besides the differences mentioned above, the method used for marking medicines is also a peculiar feature of the Chinese system. In this system bar coding is the technology adopted instead of datamatrix technology [9].

C. Medicines Traceability in Argentina.

In Argentina implementation of medicines traceability began with the publication of Resolution n.º 435 in 2011 by the Health Ministry [10], which was later detailed by "Administración Nacional de Medicamentos y Tecnología

Médica" (ANMAT) through Regulamentation n.º 3683, also in 2011 [11]. The track and trace scope follows the same line of the other nations, aiming at covering traceability of all medications commercialized in the country, by means of implementing the system in waves, the first of which has the goal of controlling traceability of products regarded as critical.

Again, in terms of control the structure adopted is similar to the one in the Turkish model. Both consist of two layers and a surveillance agency which centralizes the control of critical movements within and among organizations.

The main difference between the systems is their building. In the Argentinian system, the ANMAT formally adopted several components of the standard proposed by EPCGlobal¹. Based on this fact it might be concluded that from ANMAT's point of view the adoption of a global standard is a facilitating factor in the implementation of the system.

D. Medicines Traceability in Europe.

In Europe the implementation of medicines traceability was a public-private initiative began with a pilot in Sweden. The operational phase of the pilot lasted from September 2009 to February 2010. Its scope was reduced, for it only included 14 manufacturers, 25 products, and 25 pharmacies.

The implementation of medicines traceability in Sweden was the result of a joint effort made by the *European Federation of Pharmaceutical Industries and Associations* (EFPIA), the *European Association of Euro-Pharmaceutical Companies* (EAEPC), the *Groupement International de la Repartition Pharmaceutique* (GIRP), and the *Pharmaceutical Group of the European Union* (PGEU) [12].

The aim of the project in Sweden was to gather information to support the development of regulations on medicines traceability in the European Union (EU). Based on that the *European Falsified Medicines Directive* (EU-FMD) was published. It defines December 2017 as the final deadline for the implementation of the traceability system to be concluded. A further development of this effort was the creation of the *European Medicines Verification Organisation* (EMVO), a non-profit organization, kept by the medication registration holders, with the goal of providing the system, called *European Medicines Verification System* (EMVS), for tracking medicines in the EU member states [13].

The control model of the EMVS system consists of three layers: the corporate layer, the national layer, and the regional layer. The corporate layer is responsible for controlling the traceability events which take place within organizations. A regional HUB is fed by data from the registration holders, answering for the distribution of data about medicines to the national layer. The national layer in its turn is responsible for answering the information queries

of pharmacies and distributors.

It is important to point out that in the EMVS verification model of the medicine serial number in its regional layer is mandatory for the pharmacies prior to sale of the medicine and optional for the distributor.

Two characteristics that differentiate the European model from the others are:

- a. the optional query of the medicine traceability in the distributor, and
- b. the transfer of responsibility for traceability to a non-profit organization.

The non-mandatory feature of traceability in the distributor does not imply a greater difficulty of relating problems with medicine quality due to deviations from the good practices of medication storage and distribution. And the participation of a non-profit organization in the process may mean that, in comparison with other models, the traceability here could be implemented with less need of public investment.

However, in spite of the fact that EMVO is a non-profit organization, in case the estimate made as to the prices charged for similar services by private companies in the United States and Brazil is really confirmed, the costs of association to the organization are significantly higher. This might eventually make it impossible for small manufacturers to become associated in EMVO, what can be actually understood as a deviation of its purpose in the sense of creating a commercial barrier to newcomers to the European market.

E. Medicines Traceability in the United States.

In the United States medicines traceability is a long-time demand, which began in California in 2004, when the bases for the concept of e-pedigree were established in order to prove authenticity of medicines by means of electronic documents concerning their movements in the chain [14].

In 2013 when the *Drug Quality and Security Act* (Law 113 – 54) was published, it was defined that medicines traceability would have national reach and that, similar to what had happened in other nations, the implementation process would happen in three waves [15]:

- a. In the first wave, with deadline set for 2015, traceability will occur at batch level, with verification and storage of movements;
- b. In the second wave, with its beginning set for 2017 and its end in 2020, traceability will come to occur at the level of serialized items and boxes, again with verification and storage of movements;
- c. In the third wave, set for 2023, traceability will become mandatory at item level.

The control structure of the traceability system in the United States, unlike what happens in the other countries analyzed, has only one layer. In the American model only the organizations that participate in the pharmaceutical chain are responsible for building medicines traceability.

¹ www.g1.com

The logic used in the model is one of transferring the data necessary to building a track and trace point to point system among the organizations and towards the patient. From this perspective, Fig. 2 presents the physical and logical flow of serialized medicines distribution in the USA.

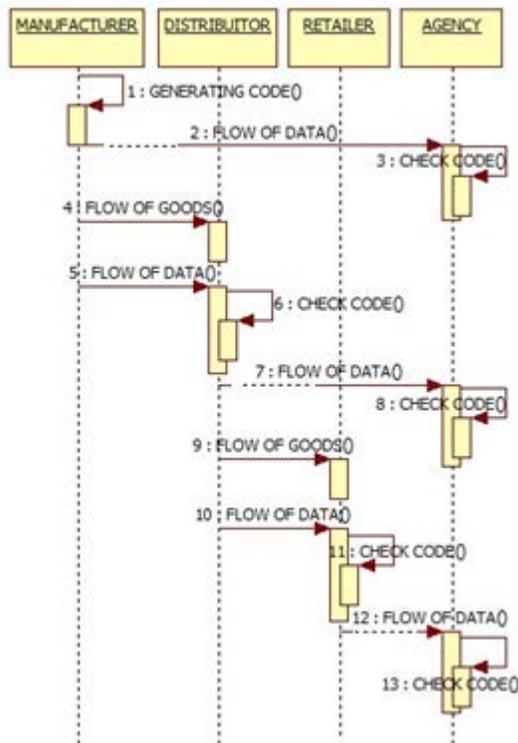


Fig. 2: Physical and logical flow of serialized medicines distribution in the United States.

With the point to point transfer it is possible to recover the whole record of movements that took the medicine to the observation point, in any organization belonging to the chain. Furthermore, in this model terms, the agency can check such a record at any time.

The American model strategy has a progressive feature of improvement of control over the authenticity of medications. In the first phase of implementation the consumer is able to retrieve data at batch level, i.e., retrieve data which confirm that a certain batch was sent to the link of the chain where the query is being made. In the second phase the consumer is able to confirm that the batch was sent to the link where the query is being made, as well as confirm that the serial number of the medication belongs to the set of numbers associated to the batch under scrutiny. In the third phase the consumer can check that the batch was sent to the link, that the serial number belongs to the set associated with that batch, and that it is expected that the serial number under scrutiny is in the link where its authenticity is being investigated.

Another feature that differentiates the American strategy from those of other countries is that it allows the organizations themselves freedom to define the best way of establishing communications control among the links of the

chain. This flexibility of the model, which had already been recommended in the Californian law in 2004, brought about new elements to be considered in designing a track and trace solution [16] [17] [18]. Within the scope of this paper four aspects are pointed out:

- Volume of data: The concern with the volume of data stored and in transit among the links;
- Technological partnership: The feasibility of having a technology partner responsible for operating data exchange among the links;
- Governance: The definition of roles and responsibilities of organizations (manufacturers, distributors, retailers, technology partners, and surveillance agency) in view of flaws in the building of track and trace data;
- Standardization: The standardization of the architecture of systems, processes, data models, message layouts, and other elements necessary to building a track and trace system solution.

III. MEDICINES TRACEABILITY IN BRAZIL

The implementation of SNCM will happen in two phases. In the first, called pilot, the holder will have to prove the efficiency of his track and trace systems in three batches of medicines, from manufacturing to dispensation point. This phase is expected to end in 2015. In the second phase, which is supposed to end in 2016, the holder will have to expand the capacity of his systems to track all his products.

In Brazil, similarly to what happens in the United States, the organizations which take part in the pharmaceutical chain are also responsible for the control of the SNCM. However, aiming to fulfill the goal set by the Brazilian State of employing medicines traceability to fight the use of medicines whose source is not certified, the commercialization of medicines with no therapeutical equivalence, and to eliminate transportation and storage practices that endanger the quality of medications, the Brazilian model differs from the American in the way control is made.

In this aspect, the basic difference of the Brazilian model is that the medicine registration holder, manufacturer or importer, plays a relevant role in the process, answering to ANVISA for:

- controlling the movements of medicines within their operations;
- controlling the movements of medicines among the organizations that participate in the pharmaceutical chain;
- replying to ANVISA's queries concerning the current positions of medicines commercialized by the company, as well as the sequence of movements which took the medicines to the observation position;
- monitoring the chain and notifying deviations in the process.

For this to happen a point to point communications

connection between the registration holder and the other links in the chain is supposed to exist, making the registration holder the center of track and trace control of his medications. Fig. 3 presents the physical and logical flow of serialized medicines distribution in Brazil.

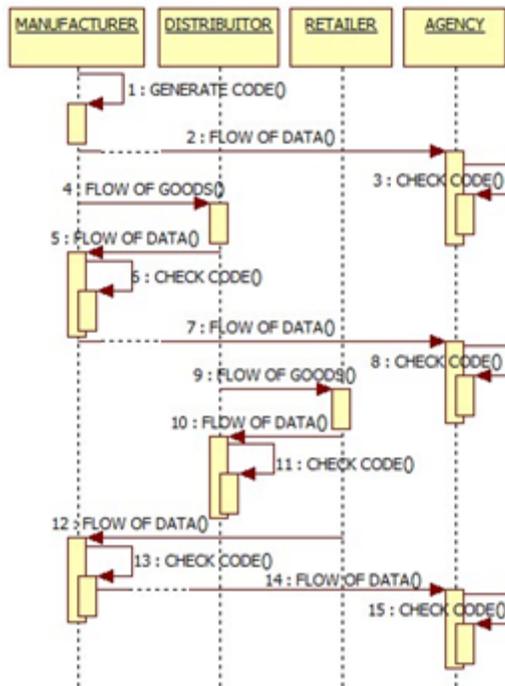


Fig. 3: Physical and logical flow of serialized medicines distribution in Brazil.

In practice the SNCM nominates the registration holder leader in the process. Thus it is natural to expect the registration holder to be even more selective when it comes to choosing his direct and indirect business partners. Giving priority to better structured partners theoretically means having partners more capable of following good manufacturing practices (in terms of storage and distribution of medicines) and less subject to improper commercial proposals, such as bonus sales.

Considering the number of agents in the pharmaceutical chain, a point to point connection among partners is not technically feasible. To solve this problem, there is the understanding that having a technological partner is essential in order to mediate communications among the links. This partner would have to:

- execute transactions of data transfer about possession of the medications among the organizations;
- make the exchange of messages generated in different standards possible ;
- validate the effectiveness of transactions;
- storage track and trace data about the medications;
- report queries of the agencies.

Based on the convergence of issues raised during the analysis of the global models studied in this paper, we have prepared a set of points about the SNCM implementation,

organized into five dimensions which can influence the implementation itself:

- Business model: The adoption of a technology partner is practically mandatory in the Brazilian model. Today the business model of technology partners consists of service fees based on the volume of medicines commercialized by the registration holders. These partners should be approved by private associations, in order to assure they have technical and financial conditions. The business model of technology partners, the authority of associations to give their approval to them, as well as the criteria for approval of technology companies, are not consensual among the participants of the chain who fear that such an approval process might result in implementation of a commercial barrier to new companies wishing to take part in the pharmaceutical chain.
- Confidentiality: One of the factors that can jeopardize the implementation of the system is the obligation that the downstream links of the chain share data with the registration holder. Sharing such data has an impact on the usual information asymmetry among the links in the chain. Information about storage levels might be used in commercial dealings in a way that is disadvantageous to the downstream links. Nevertheless, this is not the only negative factor. The level of data sharing with the technology suppliers is also sensitive for the organizations, which can resist doing that since the partner is able to store the record of several commercial transactions among manufacturers and other links of the chain, and this information can be liable to leaking.
- Standardization: Traceability of medicines is a global demand, and several nations tend to adopt open standards like the one of GS1. In cases in which no government recommendation exists, as in the USA, the market must come to an agreement. Coming to an agreement can take long and prove to be even impossible. That results, therefore, in at least three scenarios:
 - a delay in the implementation of the SNCM;
 - an implementation based on the convergence of organizations to a market standard that is established in view of its efficiency;
 - a revision of position by ANVISA, in relation to defining a standard.
- Operational efficiency: The implementation of the SNCM implies significant changes in operation. It increases the number of plant-floor, storage, and dispensation points tasks, something which might demand a manpower

increase in operation. Because of that, selection and training processes might be necessary, something which may also tamper with deadlines for compliance with the regulations.

- e. Technical feasibility: The uniqueness of the Brazilian model demands different levels of development and customization, even when one considers solutions used in other countries as platforms, because prior solutions can prove useless when SNCM implementation reaches full operational level.

IV. PROPOSAL OF SNCM PILOT PROJECT FOR HOSPITAL DAS CLINICAS - SCOPE

Considering all the open questions related to SNCM, it is necessary, in order to guarantee patient's security, to verify the effectiveness and efficiency of implementation of a track and trace solution from the point of view of operation, processes, and technology. In these terms, the Hospital das Clínicas (HC) pilot project has as its objective to map out the risks for the patient and the institution, besides preparing the organization to react to those risks.

The relevance and importance of HC for such a purpose derives from its size, caring for 5 million patients a year, with a volume of 100 million reais in purchases of medicines delivered by over 200 suppliers.

Building data about medicines traceability in Hospital das Clínicas (HC) is an operation that begins in its distribution center (DC). The center is responsible for the distribution of about 1,200 medicines, provided by a chain of more than 200 suppliers. These medicines are distributed to 11 institutes, which have approximately 250 points of dispensation. Fig. 4 presents the main stages in the distribution logistics inside the DC and HC, together with the several events of interest for the SNCM, which must be collected in each of the stages of the medicines movements.

In terms of operation it is proposed that the pilot control medicines traceability of ten business partners within the center of distribution of the Hospital, as well as their distribution in one of its institutes. This approach will allow observation of some operational difficulties in the handling of the medicines from different suppliers in the center of distribution, as well as the dispensation of medicines in different types of dispensation points in one of the institutes of HC.

In terms of processes the pilot will make possible a better understanding of the impacted processes, and also of the needs to create new mechanisms to make the gradual introduction of medicines, institutes, and dispensation points feasible.

From the point of view of technology the pilot with ten business partners will allow verification of the challenges of communication with several business partners, possibly involving the need of communication with several technology partners.

As HC is a hospital with high medication consumption,

the pilot will also contribute for the registration holders participating in it at one site to assess the readiness of their processes and systems to control the collection of events of dispensation of medicines, in compliance with IN6/2014, at a high volume.

Carrying out a pilot project as comprehensive as the one proposed means having the potential to offer a significant tool for the surveillance agency, which can use the amount of data from the pilot as a basis to validate premises assumed in regulation, as well as eventually provoke the reassessment of requirements of RDC54/2013, IN6/2014, and IT1/2015, before the beginning of the second phase of SNCM implementation, as stated by RDC54/2013.

Execution of the pilot project will allow:

- a. a clear understanding of the risks for patient and institution,
- b. definition of probability of effectiveness of risk, associated severity, reinforcement or development of mitigation mechanisms, as well as
- c. formulation and selection of proposals of reaction to risks.

Thus, carrying out a pilot project in HC is an essential step for the compliance of the institution with the regulation and very possibly a relevant contribution for the agency and the society in terms of an empirical experiment.

V. CONCLUSIONS

When presenting in this paper a proposal of scope for a pilot project of implementation of the National Medicines Control System in Hospital das Clínicas, we have considered several important aspects. First, the history of the Brazilian regulation with the challenges it presents to the largest hospital complex in Latin America. Second, the pioneer experiences of track and trace implementation in other countries and how they compare with one another. Third, we have shown the characteristics of the Brazilian law which regulates medicines traceability and, in view of implementation experiences of similar laws in the countries discussed and compared in the previous section, we have raised some still controversial issues, whose impacts on the effectiveness of the system must be taken into account. Finally, the pilot project for Hospital das Clínicas aims at implementing a medicines track and trace system which allows observation and comprehension of possible risks for patients and institution, resulting from flaws in the system.

The implementation of a project such as the one here presented benefits all parties involved in the process. As it implements medicines traceability in one part of its operations, Hospital das Clínicas will take an important step in the compliance with the law that regulates SNCM. Moreover, it is essential to underline the need for such a

pilot so that there is a mensuration of the operational risks involved, as well as a safe evaluation of the practical feasibility of the terms demanded by the law. The more empirical results are obtained in the scale of this proposal, the more feasible will it be to evaluate and improve the system. Considering the direct and indirect motives for the implementation of SNCM, there is no denying that it is in society's best interest that the system should become operational as soon as possible. On the other hand, the

impact of a non-adequate implementation of SNCM can be drastic, and this reinforces the importance that studies such as this one be carried out in multiple links of the chain, employing different perspectives, with scientific rigor, seeking a solid level of knowledge in phase with the timetable of SNCM implementation.

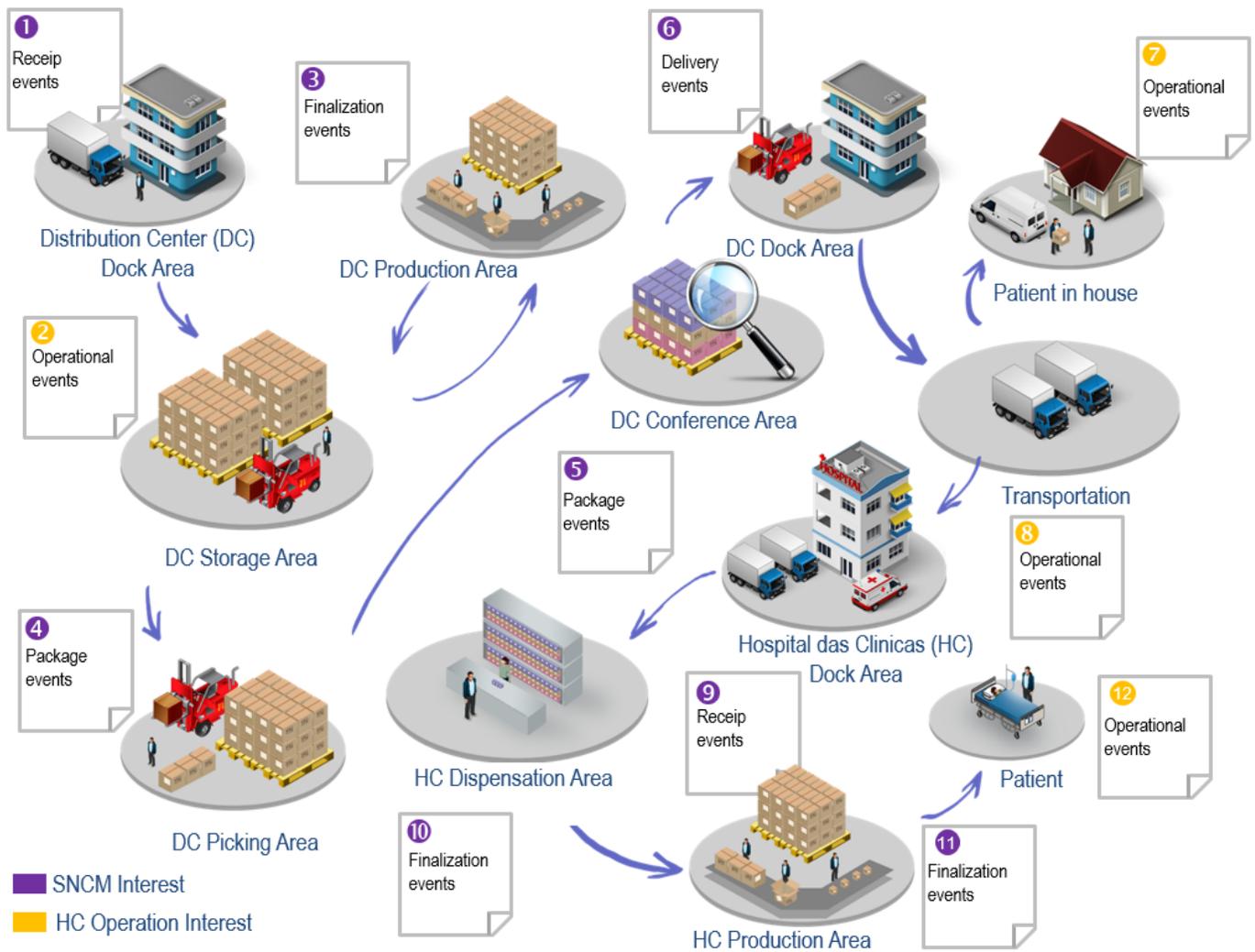


Fig. 4: SNCM Process Flow of Events inside HC

REFERENCES

- [1] *Lei n.º 11.903*, Presidência da República, 2009. Diário Oficial da União. Source: <http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2009/Lei/L11903.htm> Accessed on October, 2014.
- [2] *Reunião de Diretoria Colegiada n.º 54*, Agência Nacional de Vigilância Sanitária, 2013. Diário Oficial da União. Source: <<http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=11/12/2013&jornal=1&pagina=76&totalArquivos=168>> Accessed on October, 2014.
- [3] *Instrução Normativa n.º 6*, Agência Nacional de Vigilância Sanitária, 2014. Diário Oficial da União. Source: <<http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=19/08/2014&jornal=1&pagina=42&totalArquivos=84>> Accessed on October, 2014.
- [4] *Nota Técnica n.º 1*, Agência Nacional de Vigilância Sanitária, 2015. Diário Oficial da União. Source: <http://portal.anvisa.gov.br/wps/wcm/connect/10fc0a004800d546afd3afbd15bfe28/Nota+T%C3%A9cnica+n_1_2015_SNCM.pdf?MOD=AJPERES> Accessed on April, 2015.
- [5] *Relatório Final – Sistema de Eletrônico de Rastreamento e Autenticidade de Medicamentos*, Instituto Brasileiro de Ética Concorrencial, 2009. Source: <http://www.etco.org.br/user_file/etco-medicamentos-out2009.pdf> Accessed on October, 2014.
- [6] Calixto, J., Dias, M. L., Pokorny, M. S., & Dias, E. M. *The role of traceability in the pharmaceutical safety supply chain. Latest Trends on Systems - Volume II*, Europment, 2013. Source: <<http://www.europment.org/library/2014/santorini/bypaper/SYSTEMS/SYST EMS2-53.pdf>> Accessed on October, 2014.
- [7] Collucci, C., Cancian, C.; *Sistema para rastrear medicamentos pode atrasar uma década*. Jornal Folha de São Paulo, 2015. Source: <<http://www1.folha.uol.com.br/equilibriosaude/2015/04/1615198-sistema-para-rastrear-medicamentos-pode-atrasar-uma-decada.shtml>> Accessed on April, 2015.
- [8] Altunkan, S. M. et al. *Turkish pharmaceuticals track & trace system*. In: Health Informatics and Bioinformatics (HIBIT), 2012 7th International Symposium on. IEEE, 2012. p. 24-30. Source: <http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6209037&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxppls%2Fabs_all.jsp%3Farnumber%3D6209037> Accessed on October, 2014.
- [9] K., Guo, X., Li, M., Xie, B., Zhang, T., *Emergence of an Open Information Infrastructure in China's Pharmaceutical Distribution Industry*, 2014. Source: <<http://www.springer.com/business+%26+management/business+information+systems/book/978-3-319-05097-3>> Accessed on October, 2014.
- [10] *Resolução n.º 435*, Ministerio de Salud, 2011. Source: <http://www.anmat.gov.ar/trazabilidad/principal_en.asp> Accessed on October, 2014.
- [11] *Disposición n.º 3683*, Administración Nacional de Medicamentos y Tecnología Médica (ANMAT), 2011. Source: <http://www.anmat.gov.ar/boletin_anmat/mayo_2011/Dispo_3683-11.pdf> Accessed on October, 2014.
- [12] *EFPIA Product Verification Project – Join Report* (EFPIA), 2010. Source: <<http://www.ipha.ie/GetAttachment.aspx?id=ce8e5d3a-ca16-4d4a-8447-8c40c1d18f6d>> Accessed on October, 2014.
- [13] *EMVS Manufacturer Readiness Guidance* (EFPIA), 2013. Source: <http://www.esm-system.eu/uploads/pics/EMVS_Manufacturer_Readiness_Guidance_V1.0.pdf> Accessed on October, 2014.
- [14] *Senate Bill n.º 1307*, United State Senate, 2004. Source: <http://www.leginfo.ca.gov/pub/07-08/bill/sen/sb_1301-1350/sb_1307_cfa_20080820_175949_sen_floor.html> Accessed on October, 2014.
- [15] *Law 113-54*, United State Governnt, 2013. Source: <<http://www.gpo.gov/fdsys/pkg/PLAW-113publ54/pdf/PLAW-113publ54.pdf>> Accessed on October, 2014.
- [16] *IMPLEMENTATION GUIDELINE, Applying GSI Standards to U.S. Pharmaceutical Supply Chain Business Processes FOR THE DRUG SUPPLY CHAIN SECURITY ACT AND TRACEABILITY R1.1*, GSI, 2014 Source: <<http://www.gsi.us.org/industries/healthcare/gsi-healthcare-us/dscsa/implementation-guide>> Accessed on October, 2014.
- [17] Solanki, M., Brewster, C. *Consuming Linked data in Supply Chains: Enabling data visibility via Linked Pedigrees*. In COLD. Source: <<http://windermere.aston.ac.uk/~monika/papers/SolankiCOLD2013.pdf>> Accessed on October, 2014.
- [18] Cleland-Huang, J. et al. *Software traceability: Trends and future directions*. In: Proc. of the 36th International Conference on Software Engineering (ICSE), Hyderabad, India. 2014. Source: <<http://re.cs.depaul.edu/papers/2014-ICSE-FOSE.pdf>> Accessed on October, 2014

A tabu search using guide trees-based neighborhood for the multiple sequence alignment problem

Tahar Mehenni

Abstract—Nowadays, current Multiple Sequence Alignment (MSA) approaches do not always provide consistent solutions. In fact, alignments become increasingly difficult when treating low similarity sequences. Tabu Search is a very useful meta-heuristic approach in solving optimization problems. For the alignment of multiple sequences, which is a NP-hard problem, we apply a tabu search algorithm improved by several neighborhood generation techniques using guide trees. The algorithm is tested with the BALiBASE benchmarking database, and experiments showed encouraging results compared to the algorithms studied in this paper.

Keywords—multiple sequence alignment, tabu search, neighborhood, guide tree.

I. INTRODUCTION

Multiple sequence alignment (MSA) is a very interesting problem in molecular biology and bioinformatics. Although the most important regions of DNA are usually conserved to ensure survival, slight changes or mutations (indels) do occur as sequences evolve. Methods such as sequence alignment are used to detect and quantify similarities between different DNA and protein sequences that may have evolved from a common ancestor.

Sequence alignment is the way of inserting dashes into sequences in order to minimize (or maximize) a specified scoring function [1], [2]. There are two classes of sequencing; pairwise sequence alignment (PwSA) and multiple sequence alignment (MSA). The latter is simply an extension of pairwise alignments that align 3 or more sequences. Both MSA and PwSA can further be categorized as global or local methods. As global methods attempt to align entire sequences, local methods only align certain regions of similarity.

The majority of multiple sequence alignment heuristics is now handled using progressive approach [3]. Progressive also known as hierarchical or tree methods, generate a multiple sequence alignment by first aligning the most similar sequences and then adding successively less related sequences or groups to the alignment until the entire query set has been incorporated into the solution. Sequence relatedness is describing by the initial tree that is based on Pair wise alignments which may include heuristic Pair wise alignment methods. Some well-known programs using progressive strategies are ClustalW [4], Muscle [5], MULTAL [6] and T-COFFEE [7]. This approach has the advantages of speed and simplicity. However, its main disadvantage is the local minimum problem, which comes from the greedy nature of the approach.

Another approach is to prune the search space of the Dynamic Programming (DP) algorithm for simultaneously aligning multiple sequences, e.g., MSA [8], [9], OMA [10] etc. Algorithms of this approach often find better quality solutions than those of the progressive approach. However, they have the drawbacks of complexity, running time and memory requirement, so they can only be applied to problems with a limited number of sequences (about 10).

The iteration-based approach is also applied to the multiple sequence alignment. Iterative alignment methods produce alignment and refine it through a series of cycles (iterations) until no further improvements can be made. It is deterministic or stochastic depending on the strategy used to improve the alignment. This approach includes iterative refinement algorithms, e.g., PRRP [11], simulated annealing [12], genetic algorithms (SAGA [13], MAGA [14]), Ant Colony [15] and Swarm Intelligence [16]. Therefore, they can evade being trapped in local minima.

In this paper, we present an iteration-based approach using tabu search features to find the global alignment of multiple sequences, where the neighbors are generated using a set of operations on the guide tree of the initial solution.

The remaining of the paper is organized as follows. In section 2, we present the related work in MSA using tabu search. Section 3, describes our algorithm. Experimental results are presented in section 4 and the study is concluded in section 5.

II. RELATED WORK

Tabu Search (TS) [17], [18] was developed by Fred Glover in 1988. It was initiated as an alternative local search algorithm addressing combinatorial optimization problems in many fields like scheduling, computer channel balancing, cluster analysis, space planning etc. Tabu search is an iterative heuristic approach that uses adaptive memory features to align multiple sequences. The adaptive memory feature, a tabu list, helps the search process to avoid local optimal solutions and explores the solution space in an efficient manner.

In [19], authors propose a tabu search algorithm for multiple sequence alignment. The algorithm implements the adaptive memory features typical of tabu searches to align multiple sequences. Both aligned and unaligned initial solutions are used as starting points for this algorithm. Aligned initial solutions are generated using Feng and Doolittles progressive alignment algorithm [20]. Unaligned initial solutions are formed by inserting a fixed number of gaps into sequences at regular intervals. The quality of an alignment is measured by the COFFEE objective function [21]. In order to move from one solution to another, the algorithm moves gaps around within a

T. Mehenni is with the Department of Computer Science, Mohamed Boudiaf University, M'sila, 28000 Algeria. e-mail: tmehenni@univ-msila.dz.

Manuscript received December 13, 2014.

single sequence and performs block moves. This tabu search uses a recency-based memory structure. Thus, after gaps are moved, the tabu list is updated to avoid cycling and getting trapped in a local solution.

[22] develops in his thesis several tabu searches that progressively align sequences. He begins by a simple tabu, called Tabu A, using Dynamic Programming (DP). Then, he proposes other modified versions of tabu search, using at each time a new feature for the previous algorithm, like subgroups alignment, intensification and diversification.

In this paper, we develop a novel tabu search algorithm, by adapting similar procedures of Tabu search developed by [22], and adding a new and efficient technique for generating neighbors using guide trees.

III. ALGORITHM OVERVIEW

We first give a general description of the tabu search components of our method (initial solution, neighborhood generation and intensification method), and then provide a summarizing pseudo-code description of the main algorithm.

Tabu search works by starting from an initial solution, and iteratively explores the neighborhood of current solution by generating the moves called neighbors. In each iteration, the neighbors are evaluated through the alignment score and the best neighbor, provided it is not in the tabu list, is selected and applied to the current solution. This produces a new current solution for the next iteration. The applied neighbor is added to the tabu list and it is not allowed for a specified number of iteration called tabu tenure.

A. Initial Solution:

The generation of an initial solution is an important step towards getting a final improved alignment. A good initial solution can effectively converge faster and hence cut the computational cost. The initial solution of the tabu search is represented by a tree that is generated using the neighbor-joining guide tree (NJ) [23], which fixes the order of the partial alignments in the progressive alignment.

The NJ method constructs guide trees by clustering the nearby sequences in a stepwise manner. In each step of the sequence clustering, it minimizes the sum of branch lengths, selecting the two nearest sequences/nodes and joining them. Next, the distance between the new node and the remaining ones is recalculated. This process is repeated until all sequences are joined to the root of the guide tree. Figure 1 gives an example of a guide tree produced by 5 sequences.

The MSA is obtained from the tree as follows: the pair of sequences on the lowest level are aligned first. Then, the entire branch containing these two sequences is aligned starting from the lowest level and progressing upward to sequences on higher levels. After the MSA is determined, the alignment is scored.

The most popular scoring scheme is the sum of all pairwise alignments score: Sum-of-Pairs Score (SP).

$$SP = \sum_{i=1}^{n-1} \sum_{j=i}^n Score(S_i, S_j) \quad (1)$$

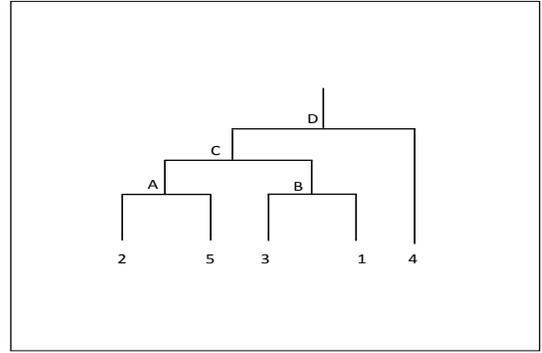


Fig. 1: An example of a guide tree generated by NJ Clustering Algorithm as Initial solution for the Tabu Search.

where

$$Score(S_i, S_j) = \max \begin{cases} (S_{i-1}, S_{j-1}) + s(x_i, y_j) \\ (S_{i-1}, S_j) - d \\ (S_{i-1}, S_j) + d \end{cases}$$

where $s(x_i, y_j)$ is the score for matching symbols x_i and y_j and d is the penalty for introducing a gap.

B. Neighborhood Generation

The neighborhood of the current solution may be generated by one of the four ways: swapping, node insertion, branch insertion or distance variation.

1) *Generation by swapping*: The simplest way of generating a neighborhood is swapping the order of the sequences (i.e. leaves) while maintaining the same guide tree topology. the number of guide trees generated by swapping is $n(n-1)/2$, where n is the number of sequences to be aligned. Figure 2 shows two guide trees (*b* and *c*) generated from the initial guide tree *a* by swapping the order of the sequences.

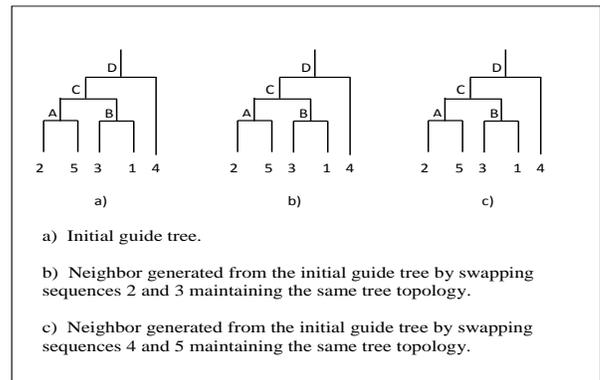


Fig. 2: Two examples of neighbors generated by swapping technique from the initial solution.

2) *Generation by node insertion:* Neighbors can be generated from the current solution (i.e. the current guide tree) by performing certain insertions of nodes. The node insertion makes it possible to move a sequence node to another location of the guide tree. This will change the topology of the initial guide tree, and the new guide tree can be considered as a neighbor of the original one.

The neighborhood can be generated randomly by this technique, since the topology of the initial guide tree is not predetermined. However, we can make only n node insertions to obtain exactly n neighbors, by selecting randomly a node to share one of the sequences (leaves) of the guide tree. More precisely, for each sequence, we choose randomly a node and move it to share this sequence, and so on. Figure 3 shows two guide trees (*b* and *c*) obtained by inserting nodes to share predetermined sequences of the initial guide tree *a*.

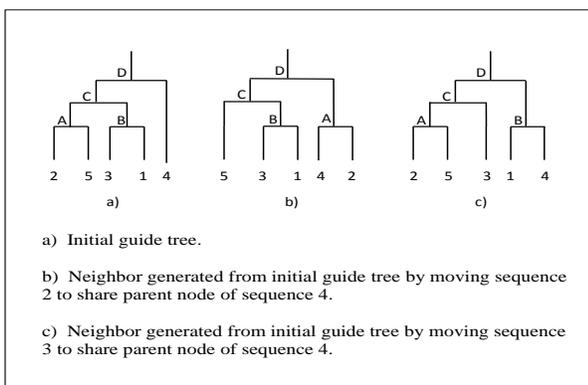


Fig. 3: Two examples of neighbors generated by node insertion technique from the initial solution.

3) *Generation by branch insertion:* Another way to generate neighbors from the current guide tree is the branch insertion, which is moving a branch of the guide tree (or a sub-tree) to another location. The new guide tree resulting of this move is considered as a neighbor of the current guide tree. This will change the topology of the initial guide tree.

Neighbors are generated randomly by branch insertion move. However, we can make only n branch insertions to generate exactly n neighbors for the current guide tree. For each sequence, we choose randomly a branch (or sub-tree) and move it to share this sequence, and so on. Figure 4 shows two guide trees (*b* and *c*) obtained by inserting branches to share predetermined sequences of the initial guide tree *a*.

4) *Generation by distance variation:* The last technique used to generate a neighborhood is the distance variation. Since the initial guide tree is obtained using NJ clustering algorithm, we can produce N different guide trees based on the NJ clustering algorithm, N being defined by the user. Each tree corresponds to a variation of the original obtained by NJ but adding some random noise into the distances in order to introduce some variability. The variation introduced in the guide tree is low enough to keep the distance criteria but significant enough to provide the necessary flexibility to

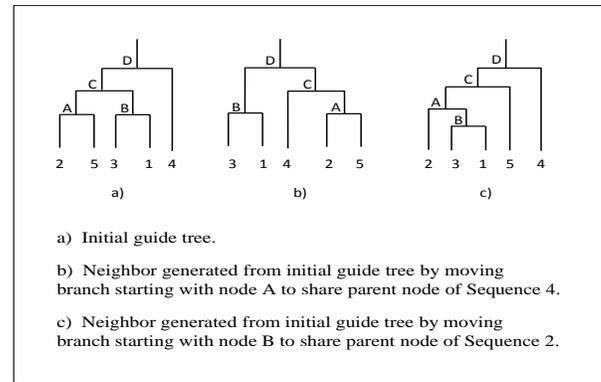


Fig. 4: Two examples of neighbors generated by branch insertion technique from the initial solution.

generate multiple alternative trees [24]. Figure 5 shows two guide trees (*b* and *c*) produced by adding variation to distances in the NJ clustering algorithm used to obtain the initial guide tree *a*.

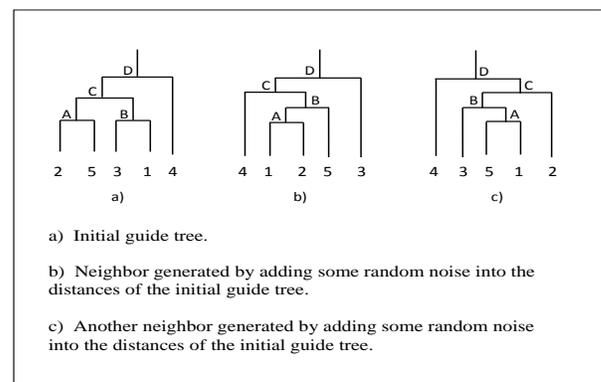


Fig. 5: Examples of neighbors generated by distance variation technique from the initial solution.

C. Intensification Method

Generally, an intensification procedure revisits and examines good solutions. It maintains the good portions of this solution and searches to find a better neighboring solution.

When a single MSA continues to have the highest score for many iterations, the intensification phase aims to escape the local minima by taking out a solution from the tabu list and restart another search process.

D. Tabu Search Algorithm

Our Tabu Search algorithm consists of generating a neighborhood of a multiple sequence σ using the techniques cited above, i.e. Swapping (SWP), Node insertion (NI), Branch insertion (BI) and Distance variation (DV). The best MSA σ'

having the higher score S_{max} is selected for the next iteration and put in the tabu list $TabuList$. This process is iterated until a T_{max} global running time is met. The pseudo-code of our tabu search algorithm is given in Algorithm 1. The details of this algorithm are explained below.

Algorithm 1 Tabu Search Algorithm for MSA

```

1: procedure GTREETABU
2:   Generate  $\sigma$  an initial MSA using NJ algorithm;
3:    $S_{max} := \text{Score}(\sigma)$ ;  $\sigma_{max} := \sigma$ ;  $TabuList := []$ ;
4:   while not  $T_{max}$  do
5:     Generate a neighborhood  $N(\sigma)$  using: SWP, NI,
     BI or DV.
6:     set  $\sigma'$  such that
7:      $S_{\sigma'} := \max_{\eta \in N(\sigma)} \text{Score}(\eta)$  and  $\sigma' \notin TabuList$ 
8:     if  $S_{\sigma'} > S_{max}$  then
9:        $S_{max} := S_{\sigma'}$ ;  $\sigma_{max} := \sigma'$ 
10:      Insert  $\sigma'$  in  $TabuList$ 
11:     end if
12:     set  $\sigma := \sigma'$ 
13:   end while
14: end procedure

```

After generating an initial solution using NJ clustering algorithm, its score is computed. While a time execution T_{max} is not reached, the tabu search is iteratively executed. Each iteration begins by generating the neighborhood of the current solution by one of the techniques among: Swapping, Node insertion, Branch insertion, Distance variation. For each neighbor, we compute its score in order to set the best neighbor having the highest score as the new current solution. This new solution is inserted in the tabu list which has a variable length depending on the number of iterations with or without improvement. If there is improvement in a certain number of continuously iterations, the length is increased in order to insert other possible solutions. The length of tabu list is decreased if within many iterations there is no improvement. In this case, a solution will be get out from the tabu list in order to restart another search process in the intensification mode.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed approach is implemented in MATLAB and tested on Intel Core i3-380M Laptop with 2 GB. To demonstrate the effectiveness of our approach, we have evaluated it on BALiBASE 2 benchmark base [25]. BALiBASE is a database of manually refined multiple sequence alignments. It can be viewed at <http://www-igbmc.u-strasbg.fr/BioInfo/BALiBASE2/index.html> or can be downloaded from <ftp://ftp-igbmc.u-strasbg.fr/pub/BALiBASE2/>.

BALiBASE database is divided into five reference sets. Reference 1 contains alignments of equidistant sequences of similar length, with no large insertions or extensions. Reference 2 aligns up to three "orphan" sequences (less than 25% identical) from reference 1 with a family of at least 15 closely related sequences. Reference 3 consists of up to 4 sub-groups, with less than 25% residue identity between sequences from different groups. The alignments are constructed by adding

TABLE I: Results given by tabu search using four neighborhood generation techniques on the BALiBASE benchmark database.

Neighborhood	Ref1	Ref2	Ref3	Ref4	Ref5	Average
SWP	90.0	93.0	76.3	87.4	85.1	86.36
NI	90.1	90.0	78.5	85.6	93.3	87.50
BI	90.05	93.8	80.7	93.7	97.9	91.23
DV	90.0	93.5	82.0	91.8	95.1	90.48

homologous family members to the more distantly related sequences in reference 1. Reference 4 contains alignments of up to 20 sequences including N/C-terminal extensions (up to 400 residues), and Reference 5 consists of alignments including internal insertions (up to 100 residues) [25].

We analyzed the tabu search results from two aspects. The very first set of tests was aimed at to verify the efficiency of our techniques of generating the neighborhood. The techniques are: Swapping (SWP), Node Insertion (NI), Branch Insertion (BI) and Distance Variation (DV). For each neighborhood technique, we ran an extensive set of tests on all the datasets provided by BALiBASE, and computed the scores. The scores using tabu search with each neighborhood generation technique are shown in Table I. The Number of Test Cases in Ref1, Ref2, Ref3, Ref4 and Ref5 are respectively 82, 23, 12, 12 and 12.

One can see in Table I that all the neighborhood generation techniques perform well in average for all the reference sets. However, it seems that Branch Insertion and Distance Variation give the best results for all the sequences of Ref2, Ref3, Ref4 and Ref5. Node insertion gives best results for sequences of Ref1. We can see that, for all the datasets provided by BALiBASE, Swapping is not the adequate neighborhood technique. This can be explained by the nature of the neighbors generated by a certain technique. For the Swapping technique, the neighbors have the same topology, so they are not very different and this will not give more amelioration of the alignment score. For the rest of techniques, the neighbors have not the same topology, but Branch insertion and Distance variation techniques seem to generate more complex guide trees, and this will give more chances to explore different solution spaces and thus, ameliorate the alignment score.

In order to verify the efficiency of our algorithm, we performed another set of tests where the results of our tabu search algorithm using a certain neighborhood technique is compared to other MSA tools. For each references set, we use the adequate neighborhood generation technique which gives the best results, and compare it to the most competitive MSA tools in the literature, such as CLUSTALW 1.83 [4], SAGA [13], MUSCLE [5], ProbCons [26], T-Coffee [7], SPEM [27], PRALINE [28], IMSA ([29] and Tabu Search developed by [19] (called in this paper TS-Riaz). Except for SAGA and TS-Riaz, which are taken from [19], the results of the other

TABLE II: Results given by Tabu Search using neighborhood techniques compared with other methods on the BALiBASE benchmark database.

Method	Ref1	Ref2	Ref3	Ref4	Ref5	Average
CLUSTALW	85.8	93.3	72.3	83.4	85.8	84.12
SAGA	82.5	95.4	77.7	78.0	86.8	84.08
MUSCLE	90.3	64.4	82.2	91.8	98.1	85.36
ProbCons	90.0	94.0	82.3	90.9	98.1	91.06
T-Coffee	86.8	93.9	76.7	92.1	94.6	88.82
SPEM	90.8	93.4	81.4	97.4	97.4	92.08
PRALINE	90.4	94.0	76.4	79.9	81.8	84.5
IMSA	83.4	92.1	78.6	73.0	83.6	82.14
TS-Riaz	76.0	88.9	71.5	77.3	90.5	80.84
TS-SWP	90.0	93.0	76.3	87.4	85.1	86.36
TS-NI	90.1	90.0	78.5	85.6	93.3	87.50
TS-BI	90.05	93.8	80.7	93.7	97.9	91.23
TS-DV	90.0	93.5	82.0	91.8	95.1	90.48

programs are taken from the work of Layeb et al. [30].

The results of our method illustrate clearly the effectiveness of using Tabu Search to perform the multiple sequence alignment. As it can be seen in Table II, our algorithm performs well in all the references sets. Our method gives good results compared to the other MSA tools. In fact, it gives the second best score for the sequences set Ref4, the third best score for Ref3 and Ref5, and it is in the fourth place for the remaining sets, i.e. Ref1 and Ref2. We can see in Table II that our Tabu search using Branch Insertion neighborhood technique has a good place for three sequences sets over five, i.e. Ref2, Ref4 and Ref5. Using the Distance Variation neighborhood technique gives the third best score for Ref3 set, and Node Insertion gives the fourth best score for Ref1. It can be seen overall, that our tabu search method using Branch Insertion neighborhood technique gives in average the second best score compared to the other algorithms studied in the paper.

V. CONCLUSION

In this paper we have demonstrated the efficiency of using tabu search to align multiple sequences. Our algorithm uses several neighborhood generation techniques. To evaluate our approach, we have used BALiBASE benchmark. Firstly, we studied different techniques to produce the neighborhood, then we compared our algorithm to the most recent and competitive MSA tools. We have observed through experiments on BALiBASE that for reference 1 and reference 2, the alignments generated by our method are encouraged. For the remaining

references, tabu search performs better than most of the other methods studied in this paper.

There are several issues for future work. First, tabu search comes with a number of parameters that can be experimented with to observe the respective effect on the search process. The parameters like tabu list size, tabu tenure, termination criteria, and neighborhood size can have a direct influence on the quality of the final alignment. Further studies are needed to test different scoring schemes and tabu search features.

REFERENCES

- [1] A. Abbas and S. Holmes, "Bioinformatics and management science: some common tools and techniques," *Operations Research*, vol. 52, no. 2, pp. 165–190, 2004.
- [2] C. Shyu, L. Sheneman, and J. Foster, "Multiple sequence alignment with evolutionary computation," *Genetic Programming and Evolvable Machines*, vol. 5, pp. 121–144, 2004.
- [3] C. Kemena and C. Notredame, "Upcoming challenges for multiple sequence alignment methods in the high-throughput era," *Bioinformatics*, vol. 25, pp. 2455–2465, 2009.
- [4] J. Thompson, D. Higgins, and T. Gibson, "ClustalW: improving the sensitivity of progressive multiple sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res.*, vol. 22, pp. 4673–4680, 1994.
- [5] R. Edgar, "MUSCLE: Multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res.*, vol. 32, pp. 1792–1797, 2004.
- [6] D. G. Higgins and W. R. Taylor, *Multiple sequence alignment, Protein Structure Prediction -Methods and Protocols*,. Humana Press, 2000.
- [7] C. Notredame, D. Higgins, and J. Heringa, "T-Coffee: a novel method for fast and accurate multiple sequence alignment," *J. Mol. Biol.*, vol. 302, pp. 205–217, 2000.
- [8] S. K. Gupta, J. D. Kececioglu, and A. A. Schaffer, "Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment," *J. Comp. Biol.*, vol. 2, no. 3, pp. 459–472, 1995.
- [9] D. Lipman, S. Altschul, and J. Kececioglu, "A tool for multiple sequence alignment," *Proc. Natl. Acad. Sci.*, vol. 86, pp. 4412–4415, 1989.
- [10] K. Reinert, J. Stoye, and T. Will, "An iterative method for faster sum-of-pairs multiple sequence alignment," *Bioinformatics*, vol. 16, pp. 808–814, 2000.
- [11] O. Gotoh, "Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments," *J. Mol. Biol.*, vol. 264, pp. 823–838, 1996.
- [12] J. Kim, S. Pramanik, and M. J. Chung, "Multiple sequence alignment using simulated annealing," *Comp. Applic. Biosci.*, vol. 10, no. 4, pp. 419–472, 1994.
- [13] C. Notredame and D. G. Higgins, "SAGA: Sequence alignment by genetic algorithm," *Nucl. Acids Res*, vol. 24, pp. 1515–1524, 1996.
- [14] T. Yokoyama, T. Watanabe, A. Taneda, and T. Shimizu, "A web server for multiple sequence alignment using genetic algorithm," *Genome Informatics*, vol. 12, pp. 382–383, 2001.
- [15] C. Blum, M. Valles, and M. Blesa, "An ant colony optimization algorithm for DNA sequencing by hybridization," *Computers and Operations Research*, vol. 38, pp. 3620–3635, 2008.
- [16] S. Lalwani, R. Kumar, and N. Gupta, "A review on particle swarm optimization variants and their applications to multiple sequence alignments," *Journal of Applied Mathematics and Bioinformatics*, vol. 3, no. 2, pp. 87–124, 2013.
- [17] F. Glover, E. Taillard, and D. de Werra, "A user's guide to tabu search," *Ann. Oper. Res.*, vol. 41, pp. 3–28, 1993.
- [18] F. Glover and M. Laguna, *Tabu Search*. Boston, USA: Kluwer Academic Publishers, 1997.

- [19] T. Riaz, Y. Wang, and K. Li, "Multiple sequence alignment using tabu search," in *Proceeding of Asia-Pacific Bioinformatics Conference (APBC2004)*, 2004, pp. 1–10.
- [20] D. Feng and R. F. Doolittle, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees," *Journal of Molecular Evolution*, vol. 24, no. 4, pp. 351–360, 1987.
- [21] C. Notredame, L. Holmes, and D. Higgins, "COFFEE: an objective function for multiple sequence alignments," *Bioinformatics*, vol. 14, no. 5, pp. 407–422, 1998.
- [22] C. Lightner, "A tabu search approach to multiple sequence alignment," Ph.D. dissertation, North Carolina State University, Raleigh, North Carolina, 2008.
- [23] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Mol. Biol. Evol.*, vol. 4, no. 4, pp. 406–425, 1987.
- [24] M. Orobítg, F. Guitaro, F. Cores, J. Lladós, and C. Notredame, "High performance computing improvements on bioinformatics consistency-based multiple sequence alignment tools," <http://dx.doi.org/10.1016/j.parco.2014.09.010>, 2014.
- [25] A. Bahr, J. D. Thompson, J. C. Thierry, and O. Poch, "BALiBASE (benchmark alignment database): enhancements for repeats, transmembrane sequences and circular permutations," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 323–326, 2001.
- [26] C. Do, M. Mahabhashyam, M. Brudno, and S. Batzoglou, "ProbCons: Probabilistic consistency-based multiple sequence alignment," *Genome Res.*, vol. 15, no. 2, pp. 330–340, 2005.
- [27] H. Zhou and Y. Zhou, "SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures," *Bioinformatics*, vol. 21, pp. 3615–3621, 2005.
- [28] V. Simossis and J. Heringa, "PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information," *Nucleic Acids Res.*, vol. 33, pp. 289–294, 2005.
- [29] V. Cutello, G. Nicosia, M. Pavone, and I. Prizzi, "Protein multiple sequence alignment by hybrid bio-inspired algorithms," *Nucleic Acids Research*, vol. 39, no. 6, pp. 1980–1990, 2010.
- [30] A. Layeb, M. Selmane, and M. Bencheikh ELhoucine, "A new greedy randomized adaptive search procedure for multiple sequence alignment," *International Journal of Bioinformatics Research and Applications*, 2011.



Tahar Mehenni Received the Engineer degree in computer science from University of Constantine, Algeria, in 1992, and the Magister degree in computer science from the University of M'sila, Algeria, in 2006. He is a PhD student in the University of Bejaia, Algeria. He has been working in the area of data mining since 2007. His research work focuses on multi-database mining and multi-relational data mining. His current research interests include Sequence Alignment, gene expression and microarrays.

He also works for a long time in software engineering, meta-heuristics, scheduling and optimization problems.

Computational automation in modern personalized medicine - AirPROM project prespective

Michal Kierzynka, Marcin Adamski, Andreas Fritz, Dmitriy Galka, Ian Jones, Dieter Maier, Andrew Wells and the AirPROM Consortium

Abstract—Modern medicine therapies tend to generate and rely on an immense amount of data that are usually produced by CT, MRI and other imaging techniques as well as genetic data coming from NGS sequencing. In order to plan a patient-specific therapy these data need to be efficiently analyzed and interpreted per individual subject. The EU-founded AirPROM project (Airway Disease Predicting Outcomes through Patient Specific Computational Modeling) is a prime example of joint cooperation that aims to develop tools to predict the progression of selected diseases and response to treatment in the area of respiratory medicine. This would not be possible without support of computer science methods. In particular, a lot of effort has been spent to integrate different software tools and present them to specialists in a form of one unified system that may be used without in depth ICT knowledge. This paper presents selected tools and techniques used to achieve this goal.

Index Terms—personalized medicine, asthma, COPD, automation, high performance computing, OpenStack cloud systems, international projects

I. INTRODUCTION

IN the recent years modern ICT technologies have tremendously improved many areas of life, including medicine-related sectors. In particular, advanced medical imaging, simulation and statistical software tools analyzing large DICOM and genomic data sets have benefited most from this development. As a result, the personalized medicine has evolved from being only a future dream to almost everyday reality. However, these advances would not be possible without a wide support from the ICT sector. Sometimes it takes a supercomputing center to analyze all the data coming from a hospital. Moreover, dedicated computational workflows are needed in order to save time by making the computations as automatic as possible, remembering that someone's health depends on them.

One important application area for these techniques is respiratory medicine. Lung diseases such as asthma and chronic obstructive pulmonary disease (COPD) affect the lives of over 500 million people worldwide [1] and costs the European Union alone more than 56 billion euros per year. Even though doctors have access to an immense amount of information and

data regarding these diseases, few new therapies have been developed [2], [3]. The AirPROM project aims to develop tools to predict the progression of disease and response to treatment for individual subjects. The project aims also at building multi-scale simulation models of the whole airway system, as a new way of characterizing asthma and COPD. Ultimately, this leads to a personalized treatment, i.e. the ability to find the best possible treatment for each patient.

In order to make things happen for a large number of patients and still keep the whole process transparent and understandable, the project is assisted by a Knowledge Management (KM) platform connected to a cloud-based OpenStack storage system, where the actual data are stored. Using the portal doctors may browse individual patient's data and start simulations with desired parameters, e.g. ANSYS LungModeller, on a high performance remote computing system using the QCG infrastructure. Moreover, some computational workflows are triggered automatically as soon as the data are available. As a result, the medical staff may concentrate more on medical aspects of the disease and spend more time on patient care. The AirPROM project also focuses on several other ICT-related aspects, e.g. multi-scale computational models of the airways [4], but this paper focuses mainly on the integration of selected software tools and their automation.

II. COMPUTATIONAL MIDDLEWARE

In order to facilitate the use of advanced high performance computing (HPC) infrastructure for the non-ICT-experts, Poznan Supercomputing and Networking Center (PSNC) has designed and developed special middleware software called QCG (previously known as QosCosGrid) [5], [6], [7]. Its main role is to create an unified access interface to different computing resources that are usually managed by various queuing systems, like TORQUE [8], Slurm [9], etc. QCG offers highly efficient mapping, execution and monitoring capabilities for different types of application scenarios, such as parameter sweep, workflows, MPI or hybrid MPI-OpenMP. The QCG middleware allows also the large-scale applications, multi-scale or complex computing models to be automatically distributed over a network of computing resources with guaranteed quality of service (QoS). Finally, the middleware also provides a number of unique features, e.g. co-allocation of distributed resources or advance reservation.

Figure 1 presents a simplified architecture of the QCG middleware. In general, it is divided into two logical levels: grid domain and administrative domain. The former, i.e. grid-level services, control, schedule and generally supervise the

M. Kierzynka is with Poznań Supercomputing and Networking Center and with Poznań University of Technology, Institute of Computing Science, Poznań, Poland, e-mail: michal.kierzynka@man.poznan.pl

M. Adamski is with Poznań Supercomputing and Networking Center, Poznań, Poland

A. Fritz is with Biomax Informatics AG, Munich, Germany

D. Galka is with Materialise NV, Kiev, Ukraine

I. Jones is with ANSYS, Inc., Oxford, UK

D. Maier is with Biomax Informatics AG, Munich, Germany

A. Wells is with ANSYS, Inc., Oxford, UK

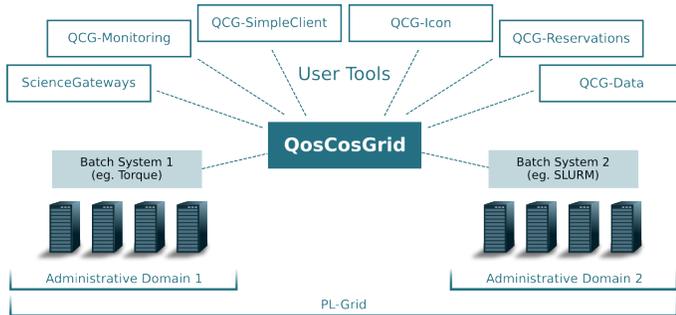


Fig. 1. A simplified architecture of the QCG middleware. Different user tools may use the same unified interface to access various computing resources that are possibly managed by different resource managers.

execution of end-users applications, which may be spread across independent administrative domains. The administrative domain represents a single HPC cluster or datacenter participating in a certain Grid or Cloud environment by sharing its computational resources. The logical separation of the administrative domains results from the fact that they are owned by separate institutions which may still control its own resource allocation and sharing policies.

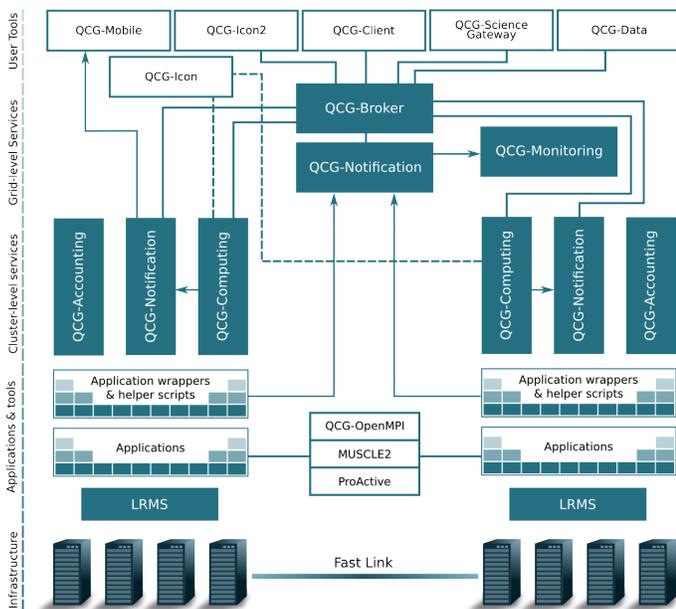


Fig. 2. A more detailed architecture of the QCG middleware illustrating the location of individual services like QCG-Computing, QCG-Notification or QCG-Broker.

Figure 2, in turn, presents a bit more detailed architecture of the QCG middleware. A key component of every administrative domain in QCG is the QCG-Computing which provides the remote access to queuing systems resources. It supports advance reservations and parallel execution environments, e.g. OpenMPI, ProActive and MUSCLE [10]. Moreover, it supports QCG Data Movement services for automatic management of input and output data transfers. QCG-Notification is another service at the administrative domain and as the name suggests is responsible for notifications, e.g. once a job has finished or failed for some reason. It is worth noting that

administrative domain services are tightly coupled with the Grid-level services. One of the key grid service is QCG-Broker which is a meta-scheduling framework controlling executions of applications on the top of queuing systems using QCG-Computing services.

The QCG middleware is used in the Polish grid infrastructure, called PL-Grid, which consists of five large supercomputing centers, namely: PSNC, Cyfronet, ICM, TASK and WCSS. Moreover, it is also used in the European Grid Infrastructure (EGI) platform and the number of its users is still growing.

The ease of use for the end-user is realized through several client interfaces, cf. Figure 2. One of them, used in the AirPROM project, is called QCG-Icon. It is a lightweight application for Windows, MAC OSX and Linux platforms, aiming to provide transparent access to applications installed on PL-Grid resources and made available via QCG services. Apart from a long list of pre-installed applications, like MATLAB, NAMD, ANSYS, Gaussian, etc., the user may easily submit, monitor and control a job defined in a form of a bash script, i.e. perform any kind of HPC computations remotely without even logging into the actual head node of the computational cluster. Since the QCG-Icon client may be parametrized, it is used in a semi-automatic way in the project. The user is presented with a choice of the computational resources, while all the patient-specific input/output data is defined automatically, which greatly reduce the burden associated with processing multiple subjects.

III. MODERN STORAGE FOR MEDICAL DATA

A. Platon U4 system

The AirPROM partners tend to generate a lot of patient-specific data, e.g. DICOM images of the lung lobes and airways, or the genetic data. Therefore, there is a need for a large, secure, reliable and fast storage system available to all the partners in the project. At the beginning of the project (early 2011) cloud-based storage systems, like OpenStack Swift, were still in their infancy, and for this reason PSNC proposed to use an archiving service called Platon U4, which was developed as a service within PL-Grid infrastructure and was already stable at that time. The main assumption was to use this reliable storage until the cloud-based systems become more mature.

The Platon U4 system [11] is a several PB (petabyte) storage system based on tapes and traditional hard disk drives which act as a cache system. Every data is replicated geographically and therefore the data is safe with respect to physical damage of parts of the system. It is also very secure due to fastidious security policy. For example, every user must be authenticated with a personal X.509 certificate and may access data only from a given location. Furthermore, all the transfers are encrypted. The data is usually accessed with SFTP protocol, using WinSCP, FileZilla or SSHFS client tools. However, the storage is also accessed by batch computations that are run on HPC machines. In this case the authentication is done via personal X.509 proxy certificate, which is generated automatically by QCG-Icon upon job submission.

The data on the storage system have a specific structure, hierarchy and naming convention. For example, every subject

has a separate directory in which every partner has its own subdirectory. Therefore, it is easy not only to browse the data manually, but also it is a convenient data structure for the software to read/write the data from/to a specific location. This is crucial from the automation perspective and will be discussed in more detail in the next sections.

B. OpenStack Swift

The notion of cloud computing has been one of the key aspects of the AirPROM project from its very beginning. An integral part of any cloud is the storage system. This plays an important role also in the AirPROM project. PSNC has been working on a cloud-based storage system for several years, and by the end of 2013 also started to introduce it to the AirPROM project. The main goal of this transition was to deliver a storage system that could fit the project requirements even better than the abovementioned Platon U4 system.

Some of the requirements for the target storage system have been defined as follows: data integrity, redundancy and high availability, role based access control and ease of integration with the applications and the Knowledge Management (KM) platform. The Platon U4 system, despite its obvious advantages, lacks some of the desired properties, e.g. the integration with the KM portal is somewhat difficult, as the user needs to have two separate accounts: one to access the portal and the other for the storage. Also different software tools need to be used when the user wants to browse information in the KM and at the same time interact with the data, e.g. download or upload DICOM images. Programmable integration of software tools with the storage was a bit complex too. For these reasons, PSNC proposed to seamlessly switch to cloud-based OpenStack Object Storage Swift software for the storage purposes.

The OpenStack Swift storage system is very different from from the Platon U4. First of all, it features a flat directory structure. Fortunately, subdirectories may be easily emulated by the "/" sign in the name of a file (here known as object). When it comes to authentication, only user name and password combination are accepted, and there is no support for X.509 certificates. Therefore PSNC has also developed additional custom authentication mechanism based on certificates. This was needed as the results of automatic computations/workflows are uploaded using proxy certificates and not with user and password credentials. Interestingly, once the user is authenticated, they receive a token which may be used to perform any storage operations without additional authentication for a given period of time, usually 24 hours. However, most importantly the OpenStack Swift may be accessed using REST API over HTTPS. This allows for relatively easy integration of the storage system with other software tools that are used in the project, possibly using one of the already existing libraries for Java, Python, Ruby or other programming languages. This means that applications may read the input data and store the results directly to the remote storage system as if they were using a local hard drive (an example is presented in Section V). Furthermore, the KM portal is now enabled to read the real content of the storage system much more easily than it was before.

IV. KNOWLEDGE MANAGEMENT PORTAL

The Knowledge Management portal is a web-based central access and data source in the AirPROM project. It integrates clinical and experimental data with information from public resources. Based on semantic mappings, the clinical world of diagnostics, physiological and laboratory attributes is connected with molecular experimental measurements and computational models which describe biological processes. Large scale data such as images are stored in the OpenStack Swift environment while the corresponding meta-information i.e. the documentation of the data as well as access rights (ACL) are managed within the KM. In addition, HPC resources connected directly to the OpenStack storage provide the horsepower for expensive computational simulations. Based on web services the OpenStack, KM and different simulation programs integrate and provide a unified experience to the user. The graphical user interface of the KM portal allows users to search for specific patients by their associated clinical as well as experimental information. The user can subsequently start imaging based simulations from web-start clients that, if needed, receive their authorization from the KM. The required data is automatically transferred from the storage system to the HPC machines for simulation. In addition, the KM controls availability of new image data on the OpenStack and automatically starts image segmentation processes on the HPC resources which results in additional information that in turn is searchable in the KM. The graphical overview of the system is presented in the Figure 3.

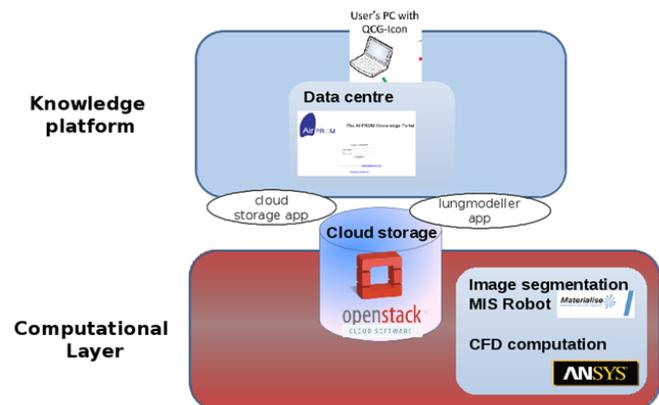


Fig. 3. A diagram presenting lung simulation environment which consists of multiple connected components: KM portal, OpenStack Swift storage, QCG managed computational resources, web-start applications and the domain specific applications.

Therefore, the KM portal acts as a central point of information about data available in the project and their semantics. With this information, the KM portal may automatically trigger computations that do not require human supervision, and the other simulations may be launched manually. However, apart from this, the KM portal is also now a central data access point in the project. Thanks to the REST API of the OpenStack Swift, the user may browse the content of the storage using the KM portal. They may also upload or download any object using a web-start application that was developed especially for

this purpose. Importantly, the application was designed to be able to resume broken uploads and was optimized for efficient retrieval of objects metadata. As a result the user does not have to use any third-party tools for data transfer, unlike at the beginning of the AirPROM project.

V. EXAMPLE APPLICATIONS

This section presents two example software tools that are used in the project and have been integrated with both the storage system and the KM portal for either automatic or semi-automatic runs in a HPC environment at PSNC.

A. Lung lobe volume computation

Materialise Mimics Robot is a server-side solution developed for unattended processing of large medical data sets. Mimics Robot can be used for automatic segmentation of 3D medical images (coming from CT, MRI, micro-CT, CBCT, 3D Ultrasound, Confocal Microscopy) as well as saving the resulting 3D models in different formats for a variety of medical and research applications. Moreover, the software is capable of measuring certain parameters of 3D models such as volume, dimensions etc. These results can be then inserted in external databases or saved locally for a later analysis. Mimics Robot is capable of supporting virtually any type of input and output storages with miscellaneous authentication methods. Storage types include but are not limited to NAS, FTP, database, OpenStack etc. The solution was designed to be scalable as it includes a built-in load balancer which takes full advantage of multi-core hardware and even multi-server network environments. The start of data processing, usually referred to as a Mimics Robot Task, can be initiated by end users via built-in Graphical User Interface or over a number of listening interfaces which can be easily integrated with external web portals. When Mimics Robot Task completes the external web portal can be notified via web-service call mechanism or, alternatively, user can be notified with an email message.

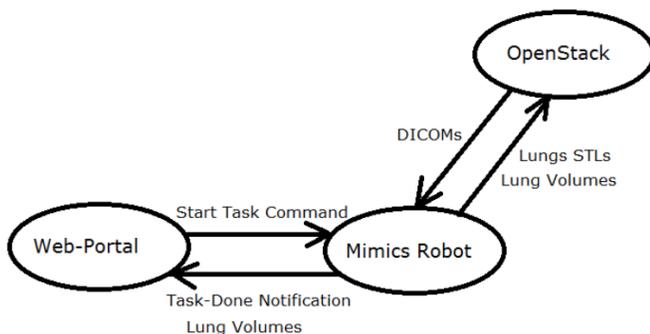


Fig. 4. An overview of the Mimics Robot workflow in the AirPROM project. The computations are started by the KM portal (Web-Portal) and the communication with the remote OpenStack storage system is done automatically. KM portal is informed once the results are uploaded to the storage and ready to use.

The Mimics Robot deployment in the AirPROM project is shown in the Figure 4. In this setup the KM portal

(Web-Portal), which is geographically located in Germany at Biomax, starts the Mimics Robot Tasks over the Internet via a web-service call. The Start Task Command identifies a study in the OpenStack storage which needs to be processed by Mimics Robot. Mimics Robot and the OpenStack storage are located in Poland at PSNC which enables relatively fast data transfer. Mimics Robot Task performs token-based OpenStack authentication, downloads relevant DICOM images, generates a Mimics project file, performs automatic lungs segmentation, measures its volumes and finally uploads resulting files to the OpenStack storage. The resulting models of lungs (in STL format) that were segmented fully automatically are shown in the Figure 5. Additionally, Mimics Robot sends a "Task-Done" notifications and lung volumes data to the KM portal via a web-service call. This way the KM portal may make these data available to interested users.

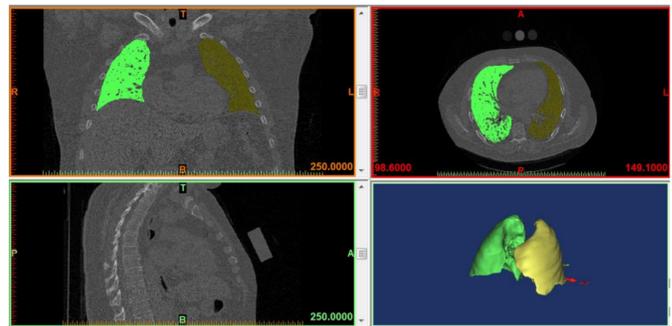


Fig. 5. The visualization of the lung lobe geometry segmented by the Mimics Robot. The left and right lung volumes are calculated correspondingly.

B. Airways resistance simulation with CFD

Within the AirPROM project, the LungModeller application has been developed as an "Extension" to the ANSYS Workbench and its CFD software, to automate the computation of the air flow in lungs, on a patient-specific basis. It has been designed from the outset to exploit cloud computing, so that the computationally expensive parts can be run using high performance computing servers, using lung geometry obtained from segmented CT scans which are stored on a centralized repository.

The user interface of the application is a stand-alone program written in Java. The application outputs a text file describing the inputs to the process, and calls a batch program to start the workflow engine. The workflow is based on the ANSYS Workbench platform [12], and is implemented using Iron Python scripts, complemented by application-specific scripts. The workflow can be run locally, or use the QCG middleware system to submit the process to the HPC cloud, and upload the relevant data files to the storage system. Alternatively, it can be initiated from the Knowledge Management system which points to a specific lung geometries located in the storage, in which case the QCG is also used to start and monitor the computations. Importantly, the communication with the storage system is performed in an automatic way and its complexity is hidden to the end user. The complete workflow

of the integrated LungModeller application is presented in Figure 6.

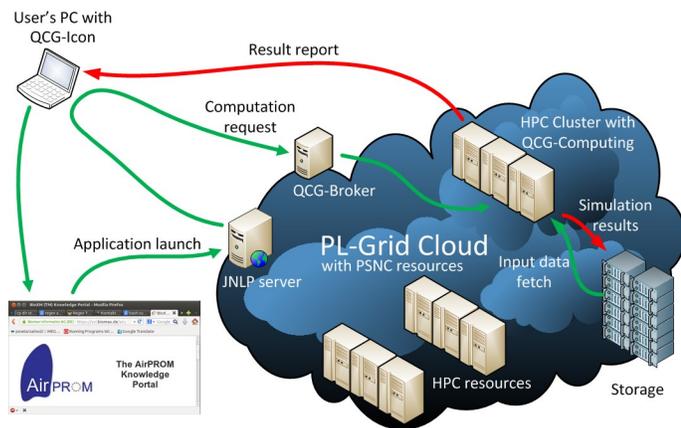


Fig. 6. A simplified diagram presenting the data flow of the LungModeller application integrated with the KM system, QCG middleware, PL-Grid infrastructure and the storage system.

The features of the CFD patient-specific workflow are as follows:

- Import of surface description meshes created by the segmentation software.
- The core of the Framework is based around the ANSYS WorkBench II workflow engine, using Iron Python as the scripting language. Future developments of ANSYS software are based around this workflow engine, and so it will be easy to update it to include new ANSYS software developments.
- Automated workflow, from GUI to reporting.
- Addition of boundary labels and boundary conditions, according to the project requirements.
- Optional creation of a skeleton, a network description of the branching structure of the airways. This is used to identify different parts of the lungs for post-processing, in particular to allow the aggregation of important clinical indicators such as flow resistance, on a lobe-by-lobe-basis. These can also be tracked in time for cases where the full transient cycle is included in the simulations.
- Volume meshing, with different options for accuracy and robustness.
- Options for using several ANSYS CFD packages, initially Fluent and CFX, to enable cross-validation of results and models.
- Automated reporting, to produce an html document with the outputs needed by clinicians and for accuracy assessment and validation.

VI. CONCLUSIONS

The joint work of several partners within the EU-founded AirPROM project has proved that the automation in modern personalized medicine is not only possible but can also greatly facilitate everyday work of medical specialists. The doctor may browse well-structured data of an individual in the KM portal to better understand cause and effect sequence for each patient. Additionally, they may analyze the patient-specific

source data using the direct connection of the KM portal with high performance OpenStack Swift storage. Moreover, the possibility of running a simulation of air flow in lungs (and possibly other software tools) in a reasonably short time using HPC environment gives additional value to the whole system. The AirPROM project may be therefore perceived as a prime example of how the computer scientists may help to develop a means to improve the overall experience of modern personalized medicine.

ACKNOWLEDGMENT

This work is funded by the EU Seventh Framework Programme FP7/2007–2013 under grant agreement no. 270194 and is presented on behalf of the whole AirPROM Consortium (www.airprom.eu).

REFERENCES

- [1] World Health Organization, *2008 World health statistics*, Geneva, Switzerland: WHO.
- [2] Hogg JC., State of the art. *Bronchiolitis in chronic obstructive pulmonary disease*, Proc. Am. Thorac. Soc. 3, 489-493, 2006
- [3] Pavord ID, Korn S, Howarth P, Bleecker E, Buhl R, Keene O. *Mepolizumab for severe eosinophilic asthma (DREAM): a multicentre, double-blind, placebo-controlled trial*, Lancet 380, 651-659, 2012
- [4] Burrowes KS, De Backer J, Smallwood R, Sterk PJ, Gut I, Wirix-Speetjens R, Siddiqui S, Owers-Bradley J, Wild J, Maier D, Brightling C, the AirPROM Consortium, *Multi-scale computational models of the airways to unravel the pathophysiological mechanisms in asthma and chronic obstructive pulmonary disease (AirPROM)* Interface Focus 3: 20120057, 2013
- [5] B. Bosak, P. Kopta, K. Kurowski, T. Piontek, M. Mamoski, *New QosCosGrid Middleware Capabilities and Its Integration with European e-Infrastructure*, In eScience on Distributed Computing Infrastructure, Springer International Publishing Switzerland, 2014, 34-53.
- [6] M. Radecki, T. Szymocha, T. Piontek, B. Bosak, M. Mamoski, P. Wolniewicz, K. Benedyczak, R. Kluszczyski, *Reservations for Compute Resources in Federated e-Infrastructure*, In eScience on Distributed Computing Infrastructure, Springer International Publishing Switzerland, 2014, 80-93.
- [7] K. Kurowski, T. Piontek, P. Kopta, M. Mamoski, B. Bosak, *Parallel Large Scale Simulations in the PL-Grid Environment*, Computational Methods in Science and Technology, Special Issue 2010, 47-56.
- [8] TORQUE Resource Manager, <http://www.adaptivecomputing.com/products/open-source/torque/>
- [9] J.B. Layton, *Caos NSA and Perceus: All-in-one Cluster Software Stack*, Linux Magazine, 5 February 2009.
- [10] J. Borgdorff, M. Mamonki, B. Bosak, K. Kurowski, M. Ben Belgacem, B. Chopard, D. Groen, P. V. Coveney, and A. G. Hoekstra, *Distributed Multiscale Computing with MUSCLE 2, the Multiscale Coupling Library and Environment*, Journal of Computational Science. 5 (2014) 719731
- [11] M. Brzezniak, N. Meyer, R. Mikoajczak, G. Jankowski, M. Jankowski, *Popular Backup/Archival Service and its Application for the Archival of the Network Traffic in the PIONIER Academic Network*, CMST Special Issue (1), 109-118, 2010.
- [12] ANSYS Workbench, <http://www.ansys.com/Products/Workflow+Technology/ANSYS+Workbench+Platform>

Intrusion detection system in area of interest using a background subtraction-based tracking algorithm

Hanbyul Chae¹, Kicheon Hong*

Abstract— In the image processing, algorithms to track the object is always difficult, yet one of the most important algorithms in the field of security. This paper deals with the system capable of automatically tracking the object within the area of interest by combining the existing tracking algorithm and another detection algorithm. In this paper, accordingly, we extract the moving object using a background subtraction technique in the video input from the camera. When the extracted object comes inside that area designated by the user, it performs automatic tracking of the object using a tracking algorithm of the Kalman Filter and Camshift. Intrusion detection and response systems can be suggested by calculating the time of its staying in areas of interest.

Keywords—Object detection, Object tracking, Kalman Filter, Camshift, Background Subtraction

I. INTRODUCTION

Nowadays in society, issues like information security and physical security in order to prepare for and guard against hacking and intrusion are spotlighted. In particular, such physical security has a lot of restrictions in the place. For example, unstable terrain such as the mountains or the sea is inevitably vulnerable to security because installation of physical security is difficult. In order to address this vulnerability, a surveillance and response system with a simple camera is installed with a relatively low restriction. And the region of interest should be automatically monitored in real time without directly monitoring and automatically tracks the attacker to be able to prepare for an emergency situation.

In [1], region of interest is monitored by the dynamic object counting program using only detection algorithm using the camera. However, the theory is dependent on the simple detection algorithm and it can't obtain a good performance if there is an object of multiple directions.

In [2], it is an algorithm to extract a person's hair and

This work was supported by the GRR program of Gyeonggi Province, Korea [(GRR SUWON 2014-B3), Development of cloud Computing-based Intelligent Video Security Surveillance System with Active Tracking Technology]. Their support is gratefully acknowledged.

Hanbyul Chae is with Dept. of Information and Telecommunications Engineering, the University of Suwon (phone: +82 1072240137; e-mail : chb423@nate.com)

Kicheon Hong is with Dept. of Information and Telecommunications Engineering, the University of Suwon (e-mail : kchong@suwon.ac.kr)

shoulder shape and to track and count objects in the specified area of interest. This algorithm shows excellent performance when a person is in the front or side but it can't detect and respond effectively to break-ins because tracking of object is difficult if someone bends or crawls. In order to detect intrusions by tracing an object, it should have outstanding detection performance and track performance notwithstanding change of shape, position and posture of an object. And even if cameras are installed at different angles, it should be able to extract and track the object continuously, which may improve the accuracy of the extracted object.

In this paper, the object is extracted [3], by using the difference between two consecutive frames, and tracked continuously using the Algorithm of Kalman Filter and Camshift when the extracted object is let into the region of interest set by the user [4]. When you acknowledge the intrusion situation based on the time change of the trajectory analysis and intrusion happens, we propose a response system which tells immediately the facts to the control room or monitoring agents.

This paper is organized as follows. Chapter 2 describes detection algorithm for moving objects through Background Subtraction, explains the detected object through a combination of Kalman Filter algorithm and Camshift algorithm and Chapter 3 describes automatic tracking and intrusion detection in the area of interest using a combination of system model, object detection and object tracking and finally Chapter 4 describes the experiments and conclusions.

II. OBJECT DETECTION USING BACKGROUND REMOVAL AND OBJECT TRACKING ALGORITHM COMBINED WITH KALMAN FILTER AND CAMSHIFT

A. Background Subtraction

Background removal is useful for extracting a foreground. There are many background removal techniques for extracting foreground but we use a very simple background removal technique. The reason for that is that the use of algorithm with more simplicity and fewer amounts of calculation can provide a system with faster processing speed rather than a complex algorithm because all you should do is to extract just moving objects. In this paper, we used the car frame technique to eliminate the background. This method is to extract the portion with the great change as much as you want as foreground after subtracting one frame from the other one. Since the image is

always affected by the noise, the difference between the pixel values less than 15 is ignored and the pixel values of large difference are displayed. That is, if the pixel value difference is small, it is set to 0 while set to 255 if it is large.

$$\delta I(x, y) = |I_t(x, y) - I_{t-1}(x, y)| \quad (1)$$

$$\delta I(x, y) = \begin{cases} 255, & \delta I(x, y) \geq 15 \\ 0, & \delta I(x, y) < 15 \end{cases} \quad (2)$$

B. Kalman Filter

Kalman Filter is a well-known method in the field of [5] motion estimation. It is a recursive algorithm for determining the best estimate that minimizes the error in the state vector in the linear dynamic system of interference by a white Gaussian noise (Linear Dynamic System). Kalman Filter is divided into three phases: prediction, measurement and modification. Effective tracking of body in Kalman Filter requires a setting of an appropriate tracking model. In this paper, we set the state vector as the center coordinates of the amount of the detected body and the change between the previous frame and the current frame, the $(\Delta X, \Delta Y)$ (x, y) .

The state vector of the Kalman Filter at time t is defined as follows.

$$x(t) = [x, y, \Delta x, \Delta y]^T \quad (3)$$

Kalman Filter is evolving along with the system state vector, $x(t)$ and the time.

$$x(t+1) = \phi(t)x(t) + \omega(t) \quad (4)$$

$\omega(t)$ is Gaussian noise whose covariance $Q(t)$ and mean is 0 and is defined as covariance $Q(t)$.

$$Q(t) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (5)$$

Measuring vector is given.

$$z(t) = H(t)x(t) + v(t) \quad (6)$$

$V(t)$ is the covariance, $R(t)$ is Gaussian noise whose covariance $R(t)$ and another mean is 0 and is defined as covariance $R(t)$.

$$R(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (7)$$

This paper is defined as a state change matrix $\phi(t)$ assuming the face is moved in a straight line with a uniform velocity.

$$\phi(t) = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (8)$$

The input vector is a four-dimensional vector of x, y coordinates and is defined as unit matrix $H(t)$ according to $\Delta x, \Delta y$ -axis and change of matrix.

$$H(t) = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (9)$$

C. Camshift algorithm

Camshift algorithm [6] It is the improved method to be used in environment of tracking Mean shift algorithm, one of color segment method and makes up for weakness of Mean shift by using method that can adjust size of search window. It is used to track the object at a high speed, and has the feature that keeps poor capability in the background of a lot of illumination change and noise. It can predict and detect the position to be changed using distribution of the Hue values of the detected object area and can track the object while it searches for the center. The RGB color model is much more sensitive to changes in illumination. Therefore, the algorithm converts the RGB color space into Hue of the HSV color model in order to reduce the impact of lighting on the track objects. It establishes a one-dimensional histogram in the region of interest, then stored and used as a tracking model.

Camshift algorithm is a variant of the meanshift algorithm already well known and is a search algorithm

operating on an empirical distribution to extract any objects. Figure 1 shows the Camshift algorithm.

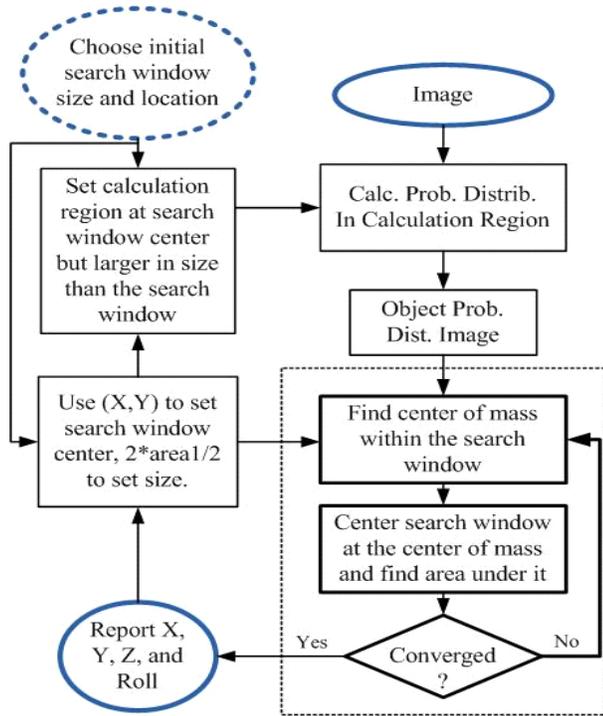


Fig. 1 Flowchart of Camshift Algorithm

III. SURVEILLANCE SYSTEM THROUGH SETTING AREA OF INTEREST AND OBJECT TRACKING

A. Structure of whole surveillance system

Surveillance system model in region of interest, which uses a camera proposed in this paper is shown in Figure 2.

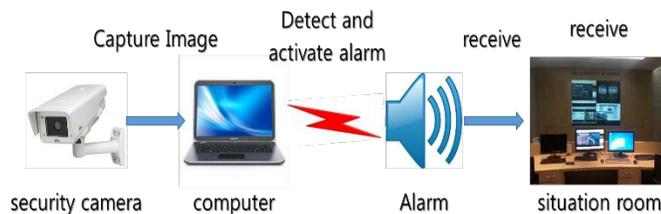


Fig. 2 Surveillance system model for area of interest using a camera

The system requires a camera and a laptop or personal computer, and the alarm signal system, which uses the precise resolution. It monitors a region of interest in real time through the algorithm which proposes a high-quality input image coming from the camera. If an intruder has occurred in areas of interest, it makes a warning alarm or reports to control center to respond quickly.

B. Object Detection Using Background Removal

It converts the video coming from a camera into a binary image by using a car frame technique to distinguish between background and foreground. It removes noise through labeling algorithm in converted binary image and extracts object. Labeling distinguishes each object by attaching a label of the same number to adjacent pixel whereas attaching a different number to components not connected. At this time, pixel sizes smaller than a certain threshold are removed in order to get rid

of the small noise.

$$Object = \begin{cases} Object, PixelSize > T \\ Noise, Otherwise \end{cases}$$

(10)

With the size of pixels in object area inspected during labeling, pixel size smaller than the threshold T is recognized as not an object but a noise.

You can see the results of car frame in Fig3 and Fig4.



Fig. 3 Input image

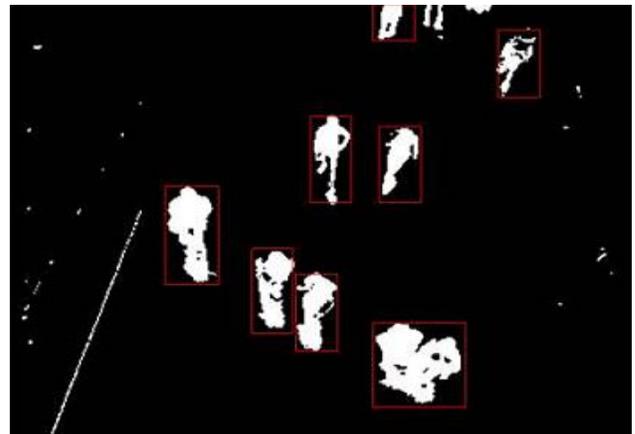


Fig. 4 Result of object detection through differential method

C. Algorithm flowchart of intrusion situation recognition

The proposed algorithm is applied to the images acquired in an indoor environment. Once one of multiple objects detected with a multi-frame method comes in area of interest designated by user, it traces the object in area of interest by applying tracking algorithm combined with Kalman Filter and Camshift[7]. Since the tracking algorithm combined with Kalman Filter and Camshift can predict the condition of object in the next frame successfully, it can trace objects properly and detect intrusion situations by analyzing the time the object stays within area of interest. Algorithm flowchart of comprehensive intrusion is as follows:

Step 1: Definition of the area of interest. The user draws a rectangle on the area he wants to monitor by setting the coordinates directly on the screen. This rectangle is the area of interest that we have set.

Step 2: Acquires images from the camera and stores it in the shape of image of RGB domain or Gray domain. Object is extracted from the differences between the current frame and the previous frame and is displayed in a rectangular framework.

Step 3: When the extracted object is let into the region of interest, it obtains the coordinates of the object within the area of interest and starts tracking through a combination of Kalman Filter and Camshift [8]. Kalman Filter makes a prediction. The prediction of the Kalman Filter is used to catch the initial position of the search window of Camshift algorithm and then Camshift performs the track.

Step 4: Set the time the tracked object stays as Time and set the threshold value as Threshold.

If, $Time > Threshold$, it considers the situation as intrusion and proceeds to Step 5.

Step 5: In case of intrusion, a warning alarm is sounded and an action is performed and the system is reset to manual mode. The overall system steps and algorithm flow chart are shown in Figure 6 below.

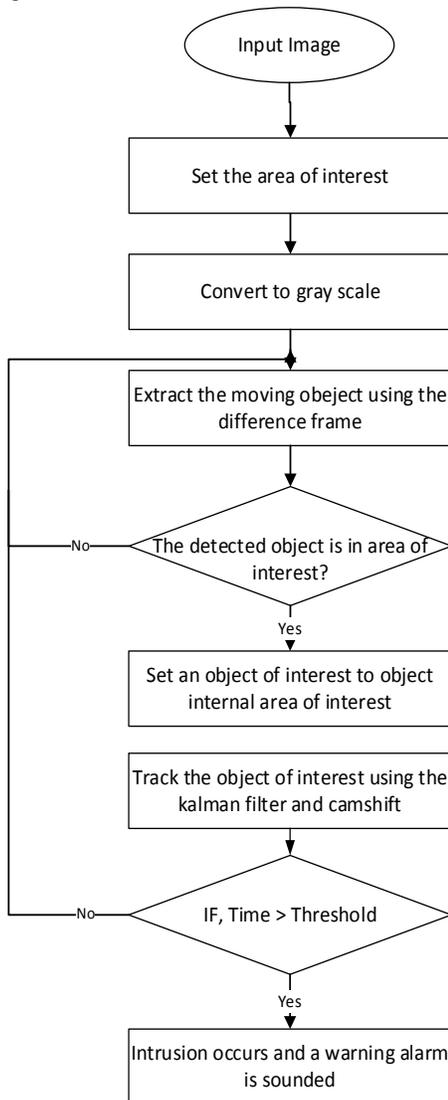


Fig. 5 Algorithm flow chart for detecting intrusion

IV. EXPERIMENT AND CONCLUSION

A. Results of Experiment on Differences between non-invasive and invasive situation

It shows the difference between results of the above-mentioned conditions and the non-intrusion condition. Images created by cutting the video frame by frame in both cases where the object extracted in the area of interest stays shorter than the threshold value or where staying longer than it, are shown in Figure 6 and Figure 7, respectively.



Fig. 6 The case where there is no intrusion within area of interest



Fig. 7. The case where there is intrusion within area of interest



Fig. 8 Notification after Intrusion detection in the area of interest.

Frames per second of the image was 29.7 and the threshold value for intrusion detection is set at 60 frames (about 2 seconds) to make an easy demonstration in this experiment. When an object comes in the area of interest, it calculates the stay time and stores it in the variable `frameup`. If the object gets out of the area of interest before `frameup` reaches the threshold, it considers the situation as non-intrusion while it does the situation as intrusion and alarm is sounded when the frame goes beyond 60. If the tracked object stays in the area of interest for more than 2 seconds, it considers the situation as intrusion and the alarm is sounded, as shown in Figure 8 whereas if it stays less than 2 seconds, it is regarded as a non-intrusion and no alarm is sounded. In order to demonstrate accuracy of the system in various environments, tests in changing angles of camera and tested in outdoor as well as indoor. Figure 9 show the images and result of the experiment where images were taken from the camera on top position of the object and Figure 10 show the result of the experiment in outdoor image.



Fig. 9 The case where camera is installed on top position of the object

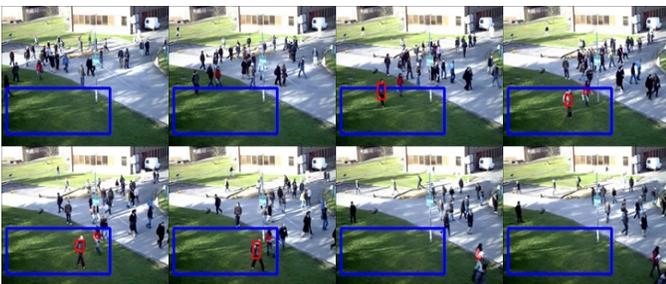


Fig. 10 The case where camera is installed in outdoor.

If an object comes within the area of interest designated by user, it will be traced continuously by the proposed tracking algorithm and time when object stays is calculated to detect the intrusion. If the object gets out of the area of interest, no more tracking is performed. Automatic tracking success rate in area of interest is proved to be high regardless of the angle of the camera.

Table 1. Intrusion detection success rate according to position of camera.

	Number of cameras	Number of successful cases for detecting intrusion	Intrusion detection success rate
<i>In case camera is installed in front of object</i>	500	442	88%
<i>In case camera is installed on top of object</i>	500	427	86%

Table 2. Intrusion detection success rate of indoor and outdoor images

	Number of cameras	Number of successful cases for detecting intrusion	Intrusion detection success rate
<i>In case camera is installed in indoor</i>	500	442	88%
<i>In case camera is installed in outdoor</i>	500	427	86%

In case the camera is positioned in front of the object, the intrusion detection success rate was found to be 88% while it was found to be 86% in position of top of the object. Automatic detection and intrusion detection rate is found to be high regardless of the position of camera. In case the image in outdoor, intrusion detection success rate is less than the image in indoor because the noise according to shadow and the movement in the wind.

B. Conclusion

This paper proposes an intrusion detection algorithm in area of interest. Object detection and object tracking results should be used to analyze and detect intrusion situations.

Dynamic object is extracted from the difference in the successive two frames of previous frame and current frame and once the extracted object comes within area of interest, it traces

the object within area of interest using automatically tracking algorithm combined with Kalman Filter and Camshift algorithm. Based on this, in this paper, it is possible to monitor in real-time the area of interest via the algorithm to extract moving object by applying background subtraction and the tracking algorithm where Kalman Filter and Camshift is combined and to trace object accurately within area of interest. Intrusion detection and response system is offered based on analysis of time-varying trajectory traced. Via the proposed system, it has shown a continuous tracking and accurate intrusion detection performance. And if an object comes into area of interest and the status object is overlapped, objects are sometimes recognized as one object while making it difficult to detect them. In order to solve it, we made tracking possible by separating them even if they come in overlapped conditions and tracing objects within any area through tracking algorithm after designating an area larger than area of interest designated initially. Later, accuracy of intrusion detection will be improved by using the shadow removal algorithm. The algorithm described in this paper is accurate in a daytime environment whereas intrusion detection is difficult because accurate detection and object tracking can't be implemented at a night environment. We are planning to configure a security system with better performance by developing an algorithm to enable precise object detection and tracking not only in the daytime but also at night and improve intrusion detection capability using thermal imaging cameras.

ACKNOWLEDGMENT

This work was supported by the GRRC program of Gyeonggi Province, Korea [(GRRC SUWON 2014-B3), Development of cloud Computing-based Intelligent Video Security Surveillance System with Active Tracking Technology]. Their support is gratefully acknowledged.

REFERENCES

- [1] Ying-Li Tian, Arun Hampapur. Automatic Counting of Interacting people by using a single uncalibrated camera. 2006 IEEE International Conference on Multimedia and Expo, pp. 1265-1268, July 2006.
- [2] Jingyu Liu, Jiazheng Liu, Mengyang Zhang. A Detection And Tracking Based Method For Real-Time People Counting, Chinese Automation Congress (CAC) 2013, pp. 470-473, Nov 2013
- [3] Gang Xu, Dong Zhao, Qi Zhou, Ding Huang. Moving target tracking based on adaptive background subtraction and improved camshaft algorithm. 2012 International conference on Audio, Language and Image Processing, pp. 919-924, July 2012.
- [4] Shengluan Huang, Jingxin Hong. Moving object tracking system based on camshaft and Kalman filter. 2011 International Conference on Consumer Electronics Communications and Networks. pp.1423-1426, 2011 April.
- [5] Kalman, R. E.A New Approach to Linear Filtering and Prediction Problems. Transaction of the ASME—Journal of Basic Engineering,82(Series D), pp.35-45, 1960
- [6] G. Bradski., Real Time Face and Object Tracking as a Component of a Perceptual User Interface. in Proc. 4th IEEE Workshop Applications of Computer Vision, pp.214-219, 1998.
- [7] Intaek Kim, Tayyab Wahab Awan, Youngsung Soh, Background Subtraction-Based Multiple Object Tracking Using Particle Filter, IWSSIP 2014, International Conference on Systems Signals and Image Processing, pp. 71-74, May 2014.
- [8] Juhyun Lee, Xiaoyu Jiang, Kicheon Hong. A Robust Faint Situation Recognition and Response System Based on Object Tracking Algorithms

Using Thermal Camera. Computer Science and its Applications Lecture Notes in Electrical Engineering, 330, pp.1233-1243, 2015.

Hanbyul Chae received the BA degree in Information and Telecommunications Engineering from Suwon University, Korea, in 2014. Hanbyul is currently a MA student at Information and Telecommunications Engineering, Suwon University, Korea.

Mr. Chae research interests include image processing, signal processing, object tracking.

Kicheon Hong received the BA in electronics engineering from Sungkyunkwan University, Seoul, Korea in 1985 and MA and Ph.D degrees in Electrical Engineering from Stevens Institute of Technology in 1988 and 1994, respectively. He was with Bell Communications Research Center (Bellcore) from 1992 to 1993 and Samsung Semiconductor Co. Ltd. From 1994 to 1998. Kicheon has been with the Department of Information and Telecommunications Engineering, The University of Suwon, Seoul, Korea, where he is currently a Professor.

Prof. Hong research interests include image processing and compression, multimedia processor, intelligent surveillance system, embedded system.

Exploratory Social Network Analysis with Pajek: Case Study on Student Group Performance

Lionel Khalil, Marie Khair, Tina Daaboul, Marie-Joelle El Hajje

Abstract— Assessing, evaluating, and predicting student's performance has always been a major research part of the academic workers aiming for academic excellence. One of the main issues related to this is identifying the major factors influencing it positively or negatively. A major related question, and not yet enough researched, is whether students taking courses together has any influence on their outcome. To answer this question, this paper presented a study of the student's behavior in course selection. We described the population considered under study, and explained the methodology followed to accomplish it. Finally, we analyzed student's performance of individual versus common courses taking several dimensions like gender, pairs or tribes, passing level (very good, good or failing). The main finding is that student cooperation in groups improves the probability of passing courses. Nevertheless, good performing students are affected slightly negatively by their contribution to the group.

Keywords—Social Network Analysis, Pajek, GPA, Student Performance.

I. INTRODUCTION

THIS paper presents a study performed on university students for observing their behavior in courses they have taken. The aim is to determine the level of performance of the students when they took their courses individually in contrast to when they took courses with their friends. Accordingly, the difference in GPA of the students in courses taken individually is analyzed and compared to that of courses taken in common. In order to reach a reliable conclusion, the relations of students between each others were mapped using Social Network Analysis and a series of steps have been followed to explore these relations, in addition to several tools such as Pajek and R, which were used to visualize the results and help us better understand the impact of the course selection on the student performance.

L. Khalil is with Notre Dame University – Louaize, Lebanon (phone: +961 9 208 118; fax: +961 9 218 771; e-mail: lkhalil@ndu.edu.lb).

M. Khair is with Notre Dame University – Louaize, Lebanon (e-mail: mkhair@ndu.edu.lb).

T. Daaboul is with Notre Dame University – Louaize, Lebanon (e-mail: tdaaboul@ndu.edu.lb).

MJ. El Hajje is with Notre Dame University – Louaize, Lebanon (e-mail: mahajje02@ndu.edu.lb).

II. LITERATURE REVIEW

Social Network Analysis SNA is the study of relationships of individuals or groups of individuals. SNA has been heavily researched and successfully tested in several fields mainly related to social sciences, human disease, scientific collaboration, business, medicine and many others. However, very little work has been done to study the SNA in the educational sector [1, 2, 3].

In general SNA involves two major directions: one direction that seeks to understand what influences the formation of relational ties in a given population, and another direction seeks to understand the impact of the relations within a SNA on a specific outcome either at an individual or at the population level [1].

When it comes to the higher educational sector, there are several disparate publications related to it. Some papers try to overlook on the subject and to show the number and the diversity of the research in it [4], others study the influence of network association on the success, or on student's research potentials, or on student integration and persistence, or even on the distribution of knowledge [1, 3, 6, 7].

In our study, a cohort of architecture students was traced. The architecture major was chosen for several reasons. First, the bachelor of architecture is very selective (out of 500 applicants only 300 are enrolled) compared to other majors (3 or 4 years). Third, the success rate within major courses is low with an average of 20% of failing students. 11% of the courses are repeated two times and 4% are repeated three times and more.

To obtain the degree of bachelor of architecture, a student must complete about 57 courses with an overall GPA of at least 2.0/4.0 and a minimum cumulative GPA of 2.3/4.0 in the Core and Major requirements.

Among the 57 courses, 43 courses are compulsory courses and 14 are elective courses. The number of sections of the same course given during the same semester is given in the Table 1 shown below. Only 16% of the compulsory courses are given in only one section. This gives the student bigger opportunity to choose among multiple sections.

In principle, students majoring in architecture tend to collaborate between each other more than other majors [8, 9]. Our hypothesis is to evaluate how much the work in groups is

affecting the whole group GPA as well as the students' individual GPA.

Number of Sections per semester	Compulsory Courses	Elective Courses
One section	16%	32%
Between 2 and 3	34%	30%
More than 4	50%	38%

Table 1 Number of Sections per Semester

III. METHODOLOGY

In order to be able to evaluate, assess, and compare the performance of the students between courses taken individually and courses taken in common, several steps were taken. The rest of this paper will cover first the procedures and steps taken to accomplish the study. Next, the treatment of the data using mainly Pajek and R in addition to other tools will be mentioned. In addition, the sample size that was taken and its validity will be presented. Finally, we will illustrate the findings and identify the conclusion.

A. Procedure and Task

Upon choosing architecture students as a population, the study will be based on evaluating whether the performance of architecture students at a private university in taking courses individually is better/worse/ or same than taking common courses with friends. Accordingly, the list of common students in architecture that have taken courses during all the six semesters inclusively from fall 2012 till spring 2015 was taken. The next step was to manipulate the population in order to get a valid sample.

B. Treatment of Data Two-mode network:

In social network analysis, matrices have been used as an efficient tool for representing a small social network and for computing results on its structure. In addition, matrices offer visual clues on the structure of small and dense networks. A matrix is a two-way table containing rows and columns. The intersection of a row and a column is called a cell of the matrix.

The selected population consisted of 311 architecture students. This data was presented using a binary matrix for each semester which shows the courses each student has taken for each semester separately. The six matrices were added to obtain a large matrix of all the courses the students have taken in the six semesters (Student-Course relation). After obtaining the matrix, R and Pajek were used for manipulating and visualizing the matrix [10]. Thus, the initial matrix is obtained which includes all the 311 student-course relation for the six consecutive semesters.

In Pajek language, affiliation networks consist of at least two sets of vertices such that affiliations connect vertices from different sets only. In the initial matrix presented (student-course relation) there are two sets, which are Students and Courses. Affiliations connect Students to Courses, not directly

Students to Students. Figure 1 shows a fragment of the Students/Courses network. This type of network is also called a two-mode network or a bipartite network, which is structurally different from the one-mode networks.

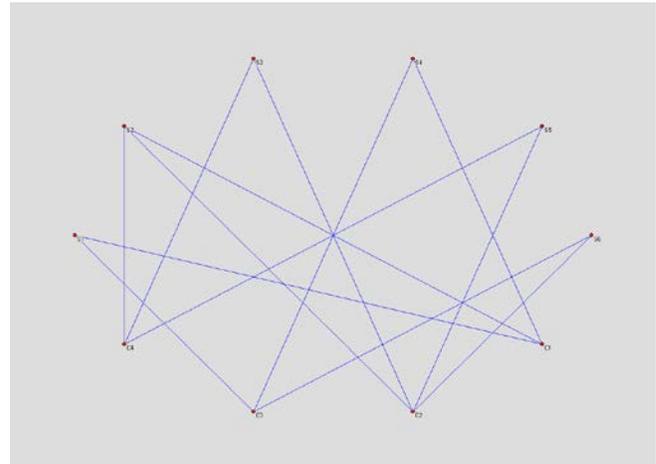


Figure 1 Students/Courses Network Example

We followed the solution commonly used to change the two-mode network into a one-mode network of students that attend common courses, which can be analyzed with standard techniques.

C. Treatment of Data: One-mode network:

Pajek has special facilities to derive a one-mode network from the two-mode network. The submenu Net>Transform>2-Mode to 1-Mode contains commands for translating two-mode into one-mode networks. A one-mode network can be created on each of the two subsets of vertices. There is an alternative way using "R" on the matrix form of the net.

By convention, vertices of the first subset Students are called rows, whereas columns refer to the second subset Courses. These terms are derived from matrix notation. Let M be the matrix having Students as rows and Courses as columns. The one-mode network for students is M^*M^T and the one-mode network for Courses is M^T*M .

In M^*M^T representing the one-mode network for students, called student-student matrix each row and column represents one vertex of the network, for instance, the first (highest) row and the first (left) column feature on Student. Social cohesion is linked to the structural concepts of density and connectedness. Density refers to the number of links between Vertices meaning how many courses students take together. A number in the cell indicates the number of courses shared with another student and a zero cell means that there is no course in common. "R" programming language was used in order to obtain the student-student matrix by calculating M^*M^T .

D. Threshold and Sample Size:

The Network has a density range from 0 up to 26 courses in common. Figure 2 shows the log-linear relation between the density and the number of vertices (students in relation).

Any Student sharing a minimum number of courses with another student is probably a friend of him or her. This hypothesis can be validated using Facebook where it was checked whether students belonging to a certain tribe or pair are really friends.

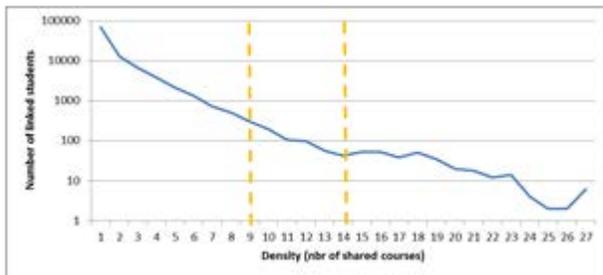


Figure 2 log representation of density per vertices

As a density threshold to deciding whether a group of students are friends or not, the number of shared courses between them was specified. Two density thresholds have been chosen, the first was 14 or more courses in common and the second was 9 or more courses in common. The density threshold 9 or more and 14 or more were chosen because students are allowed to take a maximum of 5 to 7 courses each semester. Thus, a student can take around 10 to 14 courses in a year; so students that share minimum 9 or 14 courses have been together for a minimum of a year.

Two matrices have been designed: a high density matrix with a range of density from 9 to 13 and a very high density matrix with a threshold at 14. The initial matrix has been cleaned with the diagonal value down to 0 and density relations with less than the threshold have been put at 0. The high density matrix is reduced to 94 students and the very high density matrix is reduced to 81 students. To avoid overlap between the two Networks, we first consider the tribes and pairs in the very high density network and then only the pairs from the matrix with high density that share from 9 up to 13 courses was taken into consideration. Note that one student-student relation is only shown once. This means that if this relation is shown in a very high density matrix (more than 14 common courses), it will not be shown in the high density matrix (9 to 13 common courses). In addition, if two students are part of a tribe, they will not be shown in a pair.

Hence, the high and very high density matrix (14 courses or more) were upload on Pajek to get the pairs and tribes.

E. Validity of the Sample:

The sample size required for a specified level of confidence in the result with a specified degree of sampling error is calculated based on a formula in relation to a population of a specified size [11]. But Cook et al [12] and Draugalis et al [13] point out that response representativeness can compensate a low sample. Several case studies [14,15,16] have presented low samples and identified a small effect of error indicating adequate representativeness of their sample. With 26% ratio for the sample of minimum 14 courses in common and 57% for the sample of 9% in common, the two samples fulfill and

are far above liberal conditions of 10% sampling error and 80% confidence level as per defined by Nulty [17].

F. Obtaining Tribes:

The student-student matrix was placed in Pajek that transforms it into a network where it was manipulated into groups of students i.e. showing the students sharing the same courses as tribes. There are several techniques to detect cohesive subgroups based on density and connectedness. We identified cliques or complete subnetworks based on the Pajek function Layout>Energy>Kamada-Kawai>Separate Components. Pajek divided the students into Pairs which means they share two courses together and into Tribes where they share three or more courses together.

Figure2 shows the one-mode network of students that is derived from the network in Figure 3. It is constructed in the following way. Whenever two Students share a Course in the two-mode network, there is a line between them in the one-mode network.

When students share multiple courses multiple lines are replaced by a valued single line indicating the original number of lines between two vertices, in other words the number of courses in common of two students or the density of the relation.

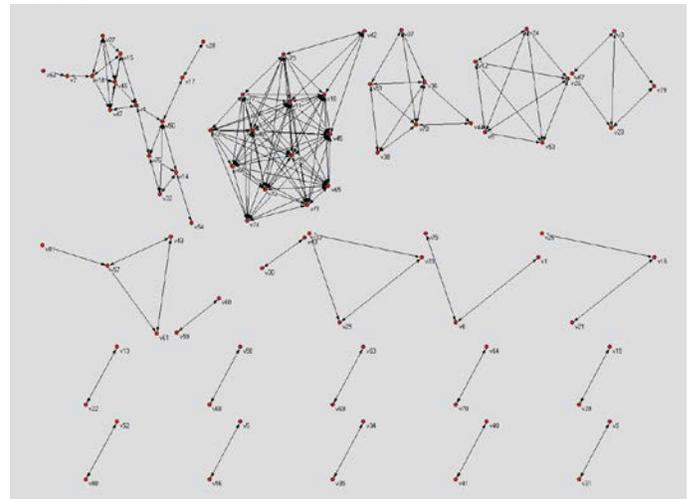


Figure 3 Tribes and Pairs in high density Network

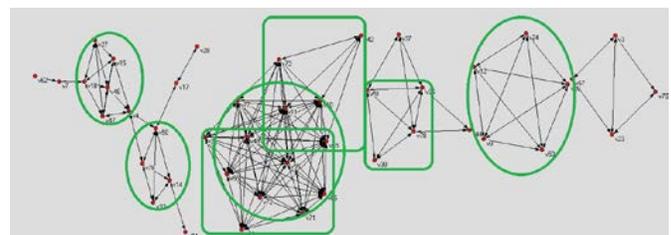


Figure 4 Cliques in the High Density network

Independent students are the ones connected with less than 14 courses (respectively 9) with other students. Pairs of students are students who are connected with bidirectional arcs with a density of more than 14 (respectively 9). Tribes are group of students who are connected with more than 14 courses two by two (respectively 9). Cliques are group of students who are all connected with more than 14 courses

(respectively 9). A clique is a set of vertices in which each vertex is directly connected to all other vertices. In Figure 4 there is only one example of clique (clique of 5 nodes); other tribes (clusters) are often cliques with additional students linked to some members of the tribes.

D. Validity of Tribes:

As mentioned above, a tribe holds set of students that have taken courses together which means that there might be a friendship relation between them. In order to make sure that our assumption is right, we have checked the friendship relation of the tribes on Facebook. Below is the result of our analysis. Table 2 shows 19 pairs out of 22 that have a 100% match with Facebook network and an overall match of 94%. Some tribes were identified on Facebook as friends from their friend list and some were identified from their common pictures. The 9 pairs denoted by “undetermined” are the ones that were either not found on Facebook or their friend list isn't viewable.

Tribe Name	Number of Stud in Tribe	FB Friends Match	Percentage of Match
Tribe 1	3	3	100%
Tribe 2	3	3	100%
Pair 1	2	2	100%
Pair 3	2	2	100%
Pair 4	2	2	100%
Pair 9	2	2	100%
Pair 10	2	2	100%
Pair 11	2	2	100%
Pair 12	2	2	100%
Tribe 4	4	4	100%
Tribe 7	6	6	100%
Tribe 9	15	15	100%
Pair 13	2	2	100%
Pair 14	2	2	100%
Pair 15	2	2	100%
Pair 16	2	2	100%
Pair 17	2	2	100%
Pair 18	2	2	100%
Tribe 8	5	4	80%
Tribe 3	3	2	67%
Tribe 5	14	9	64%
Tribe 6	4	2	50%
Percentage of Match			94%
Pair 2	2	UNDETERMINED	N.A
Pair 5	2	UNDETERMINED	N.A
Pair 6	2	UNDETERMINED	N.A
Pair 7	2	UNDETERMINED	N.A
Pair 8	2	UNDETERMINED	N.A
Pair 19	2	UNDETERMINED	N.A
Pair 20	2	UNDETERMINED	N.A
Pair 21	2	UNDETERMINED	N.A
Pair 22	2	UNDETERMINED	N.A

Table 2 Validity of Tribes and Pairs

G. Adding demographic data

Lastly, demographic data for each student in a tribe and pair was collected such as gender, grade, campus, and others. For each pair and tribe we calculated the average grade for individual and common courses, standard deviation, and percentage of courses taken. The data was also analyzed depending on gender. The following section will present the findings of the study.

IV. FINDINGS

A. Gender Inference:

1) Performance in Individual and Common Courses

According to Gender:

To begin with, the first finding concerns the performance of students regarding their gender. The goal is to identify whether female and male architecture students perform better in individual than common courses. Accordingly, the average and standard deviation of grades were calculated. The results show that females perform slightly better in individual work rather than common, having an average GPA of 2.9 and 2.86 for individual and common courses respectively. As for males the result showed that they perform better in common courses than in individual with an average GPA of 2.41 and 2.56 in individual and common respectively. The standard deviation of the GPA is slightly the same for females in both types of courses (0.98 and 0.96) and for males the standard deviation of the GPA is higher for individual courses (1.11) than in common courses (0.94).

2) Performance in Tribes According to Gender:

Another finding concerning the gender is the level of performance of students in tribes versus students in pairs according to the gender. The results show that females in tribes perform better in individual courses than females in pairs (3.09 and 2.83 for tribes and pairs respectively). As in common courses for females the average of grade is better in pairs than in tribes (2.82 and 2.95 for tribes and pairs respectively). The result for males is better in pairs (2.43) than in tribes (2.36) for the average of individual courses and the result is the same in common courses (2.56).

B. Percentage of Passing in Individual and Common Courses:

The percentage of passing and failing a course was calculated for both individual and common courses to see where the students are more successful. There are three categories of grades an architecture student can get on a course, either a passing grade (A and B range), a non-passing grade (C and D range), or a failing grade (F or withdrawal). Accordingly, the percentages of the three types of categories of grades were calculated for both individual and common courses. The results show a 79% of passing in common courses which is slightly higher than percentage of passing individual courses (75%). As for the non-passing grades the percentages are almost the same (17% and 16% for individual and common courses respectively). The percentage of failing is slightly higher in individual than common courses (8% and 5% respectively).

C. GPA Difference of Common minus Individual Courses:

Figure 5 below shows the GPA difference of common minus individual courses. The blue line represents the linear evolution of the GPA difference, the red line represents the outliers and the green line is the boundary. The results show that the average difference of GPA is negative 0.09. Figure 6 shows that regardless the percentage of common courses, GPA difference is between -0.04 and -0.02 excluding the outliers.

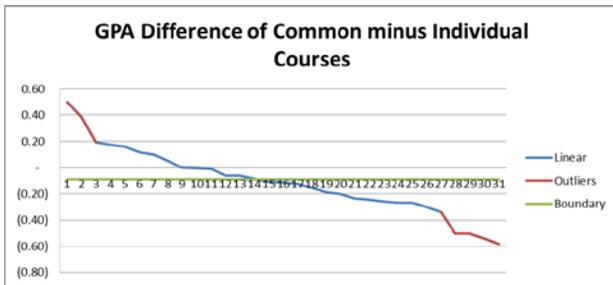


Figure 5 GPA Difference of common minus Individual Course

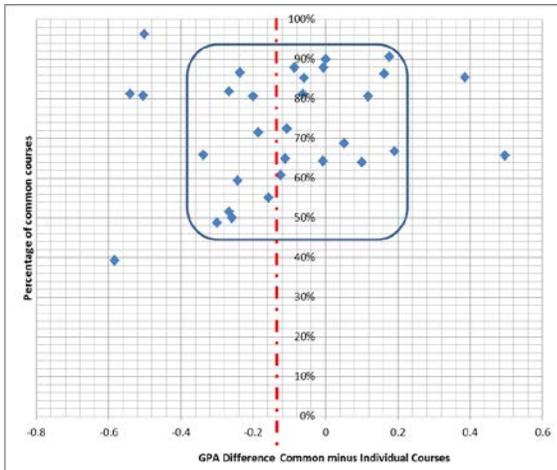


Figure 6 GPA Difference with percentage of common courses

D. Difference between Standard Deviation (Common minus Individual courses):

Figure 7 shows the difference between standard deviation of common and individual courses. Students' grades are closer to each other when they work together in 78% of cases rather than when they work alone. This result is independent of the percentage of common courses.

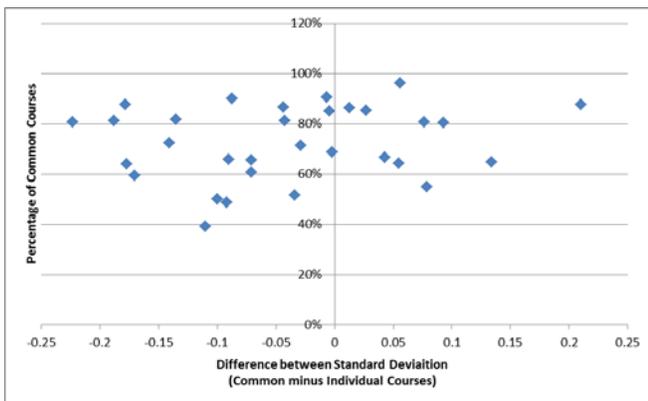


Figure 7- Difference in the Standard Deviation of GPA

V. CONCLUSION

The main conclusions which we can draw are: first that males perform better in groups while females perform the

same or worse; the standard deviation for the grades in group course is lesser than that in individual courses meaning that the students, once in groups help each other; third the number of students belonging to groups seems low (only one fourth of the students) despite the fact that they have the choice to take common courses; and finally the percentage of students passing the courses is higher in common courses than in individual courses. The main finding is that student cooperation in groups improves the probability of passing courses. Nevertheless, good performing students are affected slightly negatively by their contribution to the group.

The limitations of this study are that application of subsequent six semesters of architecture students in one university can be applied to a local context. Findings report new factors on the GPA with implications to only undergraduate students. The generalizability of findings is limited because of small sample size and area selected for sampling.

In further works, the study has to be generalized to all majors of the university. This study has not examined the decision making abilities of students within a tribe, offers an area for future research information gathered from this study and conclusions made might need further research in other university as well. The study is quantitative in nature; therefore requires further exploratory analysis in order to address remaining research questions on elements which have a significant influence on students' choice of a course.

REFERENCES

- [1] D. Grunspan, B. Wiggins, and S. Goodreau. "Understanding classrooms through social network analysis: A primer for social network analysis in education research." *CBE-Life Sciences Education* 13.2, 2014, pp. 167-178.
- [2] K. Akers, and K. Bradley. "Examining graduate committee faculty compositions-A social network analysis example.", <http://www.uky.edu/~kbrad2/Kate.pdf>, last visited May 27, 2015.
- [3] J. Hommes, et al. "Visualising the invisible: a network approach to reveal the informal social side of student learning." *Advances in Health Sciences Education* 17.5, 2012, pp. 743-757.
- [4] S. Biancani, and D. McFarland. "Social networks research in higher education." *Higher education: Handbook of theory and research*, Springer Netherlands, 2013, pp.151-215.
- [5] X. Liu, and H. Zhu. "The Influence of Friendship Network on Graduate Student's Research Potential.", in *International conference on social and technology education (ICSSTE 2015)*, 2015.
- [6] T. Scott. "Ties that bind: A social network approach to understanding student integration and persistence." *Journal of Higher Education*, 2000, pp. 591-615.
- [7] D. Rulke, and J. Galaskiewicz. "Distribution of knowledge, group network structure, and group performance." In *Management Science* 46.5, 2000, pp. 612-625.
- [8] O. Demirbas, and H. Demirkan. "Learning styles of design students and the relationship of academic performance and gender in design education." *Learning and Instruction* 17.3, 2007, pp. 345-359.
- [9] M. Mills, and C. Fullagar. "Motivation and flow: Toward an understanding of the dynamics of the relation in architecture students." *The Journal of psychology* 142.5, 2008, pp.533-556.
- [10] W. Nooy, A. Mrvar, and V. Batagelj, (2011) *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, New York
- [11] A. Astin. *Preventing Students from Dropping Out*, San Francisco: Jossey-Bass Publishers, 1975.

- [12] C. Cook, F. Heath, and R.L. Thompson . “A meta-analysis of response rates in web- or internet-based surveys.” *Educ and Psychol Meas*,2000, pp. 60(6):821–36.
- [13] J. Draugalis, S. Coons, C. Plaza, “Best Practices for Survey Research Reports: A Synopsis for Authors and Reviewers”. *Am J Pharm Educ*, 2008, 2008;(1):72. Article 11.
- [14] <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2254236/pdf/ajpe11.pdf>
- [15] L. Khalil, J. Draiby and N. Abi Karam, “Author rights awareness to promote an inter-university open-access repository for theses and memoirs” , in *SEAAIR 2014 Conference Cross-Cultural Education for AEC 2015: Realizing Possibilities, Defining Foundations*.<http://www.seairweb.info/Conference/index.aspx>
- [16] S. Sivo, C. Saunders, Q. Chang, and J. Jiang "How Low Should You Go? Low Response Rates and the Validity of Inference in IS Questionnaire Research," *Journal of the Association for Information Systems: Vol. 7: Iss. 6, Article 17. 2006.* <http://www.bus.ucf.edu/faculty/csaunders/file.axd?file=2011%2F2%2FHow+Low+Should+You+Go..Low+Response+Rates+and+the+Validity+of+Inference+in+IS+Questionnaire+Research.pdf> (accessed August 15, 2014).
- [17] J. Fincham. “Response Rates and Responsiveness for Surveys, Standards, and the Journal”, *Am J Pharm Educ*. 2008 April 15; 72(2): 43. PMID: PMC2384218, 2008, <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2384218/pdf/ajpe43.pdf> (accessed August 15, 2014).
- [18] D. Nulty . “The adequacy of response rates to online and paper surveys: what can be done?”, *Assessment & Evaluation in Higher Education* Vol. 33, No. 3, June 2008, 301–314, 2008. <http://www.uaf.edu/files/uafgov/fsadmin-nulty5-19-10.pdf> (accessed August 15, 2014).
- [19] J. Scott. *Social network analysis: A handbook*. London: Sage Publications. ISBN 0-8039-8480-4, 1991.

Warden 3: Security Event Exchange Redesign

Pavel Kácha, Michal Kostěnc, Andrea Kropáčová

Abstract—Warden is a system for efficient sharing of information about detected events (threats) from honeypots, intrusion detection systems, network threat probes and even external sources, designed as multi-client queue. Warden 3 is thorough redesign, building on previous experiences and putting together previous work on Intrusion Detection Extensible Alert (IDEA) format and up to date taxonomies, switching from SOAP to JSON flavoured HTTPS and improving filtering and authentication capabilities on the way.

Keywords—alert, security event, incident response, ids, event exchange, honeypot, json

I. INTRODUCTION

Warden is a system for efficient sharing information about detected threats, available under 3-clause BSD license. The system mimics the behaviour of the queue with multiple producers and multiple consumers – the detection probes push security events to the hub, and clients – analysers, blacklist generators, storage and aggregators can pull the new events at will. However, to take some burden out of clients and network, the server provides means for basic filtering.

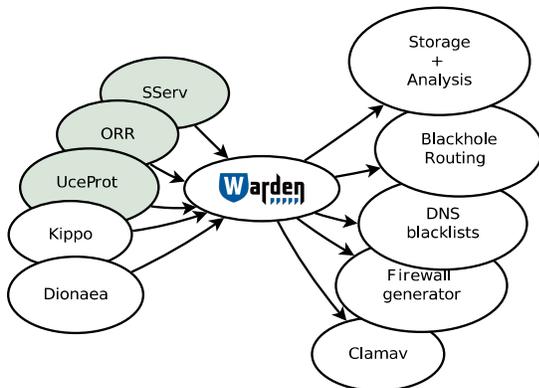


Fig. 1: Warden architecture

It is safe to say that venerable Warden 2 [11] was ambitious project and vast improvement in security incident sharing in CESNET network – the identification of new requirements and need for new directions is itself a proof that it is useful and project itself was a great achievement.

Warden 2 improved security incident handling speed in CESNET NREN and helped to reach healthier network state. Its vast body of incident data allowed for interesting data processing and correlation projects, such as [1] and [6].

This work has been supported by the CESNET association and the operator of the Czech national research and education network referred to as CESNET2 within its “Large Infrastructure” (LM2010005) research programme, running within 2010-2015 timeframe.

Pavel Kácha, Michal Kostěnc and Andrea Kropáčová work in CESNET, Zikova 4, Prague, Czech Republic (e-mails: ph@cesnet.cz, kostenc@cesnet.cz, andrea@cesnet.cz).

II. REQUIREMENTS

However, based on several years of life with previous version of Warden, real world experience and also taking into account current state of the art, let us identify possible improvements, suitable changes, and new requirements.

A. Previous Warden version review

Warden 2 used RPC calls for sending and receiving events using SOAP over HTTPS. While HTTPS shows its utility, SOAP with its complete XML (de)marshalling stack bites into performance and brings in large tree of library dependences. There is a need for lighter and more maintainable approach.

Events were represented as number of RPC call arguments (name, time, type, attack source IP and type, attack destination port, attack volume and free text note). That has shown as insufficient for many of security events nowadays in the wild (complex phishing attacks as a notable example). We need to find more flexible and extensible representation.

Also, code for validation of events must have been written by hand, specifically for our defined fields. Some standard solution would be more robust.

Warden 2 server provided basic event filtering for receiving clients, based on event type, and on simplistic notion of “own” events. This proved insufficient, users are calling for filtering based on detector type (honeypots provide greater certainty than portscan detectors, for example). Also, “own” means something different for various users, especially in organisations with complicated internal hierarchy – we need to use better representation.

Event type and detector description type was represented by loose set of categories (tags), which were added on “as needed” basis. We need to use some more standard and structured solution.

Sending API was designed to push only one event at a time, severely limiting throughput and overall sender performance.

Clients needed to save state – while this is good feature to have for complex receiving clients workflow, there is no need for simple pull-process-forget clients to be stateful.

In Warden 2, clients get authenticated by server certificate, however server certificate is usually same for the whole machine, so individual clients are differentiated only by telling its own name. However, client name is widely known, so this allows for client impersonation within one machine. We should introduce more tamper proof solution.

Last but not least – Warden 2 server was written in Perl. While it was logical choice when the project started, nowadays, when Perl 3 (and necessary ported libraries) is nowhere near to finish, number of skilled Perl programmers is on the decline, and stability and quality and compatibility of requisite libraries for Warden 2 varies wildly, even language and platform is worth reconsideration.

III. DESIGN

Lets now elaborate on stated requirements and make implementation decisions.

A. Protocols

Warden RPC calls essentially consist of parametrised event pull, unconditional event push and service calls for getting information from server. Pull can be realised by standard bare HTTP call, however as HTTP notion of return data are general “documents”, we will have to choose some way to serialise. In push direction, sending event data through HTTP GET parameters is impractical due to encoding concerns and size limits, so POST with the same serialisation format would be feasible. While we can consider XML (which is driving engine SOAP in previous Warden version), there exists much lighter solution, which gained widespread recognition, is able to directly represent fundamental data structures from various programming languages, and is often used together with various HTTP technologies – JSON.

B. Formats

The serialisation protocol closely relates to format of security events. We have already (originally for Mentat project [6]) created structured and extensible format – IDEA [3], which already uses JSON as main representation.

With IDEA we also get mature incident categorization (based on MkII [10]) and expressive set of detector description tags for free [4].

Also, there are already tools in place for IDEA, which provide validation according to JSON schema definition, solving yet another requirement.

C. Filtering

Based on experience, we don't need overly complex filtering, Warden should serve mostly as reliable transport mechanism, not data-mining store or security event search engine. IDEA gives us notion of categories and detector tags, so we will allow for positive (“has category”) and negative (“has not category”) filters on these fields. That will satisfy both use for searching by type (all events about portscan will have category Recon.Scanning) and for searching by detector type (if we want to get only events with high probability, that perpetrator really meant it, we can filter out only detectors, based on successful attack – by for example “Honeybot” tag).

D. Organisational hierarchy filtering

We are still facing problem with notion of “own” events. Two administrators from one organisation may have their own reasons to either accept each others detectors data as “own” (they already have the data internally), or to understand each others detectors data as foreign (they want the data to arrive through Warden). As we cannot force any kind of rigid resolution onto them, we have to provide solution, which allows to project their notion onto the system and use it for filtering of “own” wanted/unwanted events.

Logical solution would be using hierarchy of DNS names. However, keeping more complicated structure in DNS servers gets inconvenient very quickly, and also may unwillingly

disclose addresses of the detectors or honeypots. As we do not need complete distributed name infrastructure, we can use hierarchical IDEA Node names as the base and allow organisations to define identifiers inner structure themselves. So, modelled after Java class names, client name is dot separated list of labels, with significance from left to right – leftmost denoting largest containing realm, rightmost denoting single entity. So if we have name realm scheme akin to “org.example.csirt.honey2”, we can allow to filter based on prefixes and it is than completely responsibility of organisation, what hierarchy and names it will use and how it will filter incoming events.

We can also allow both positive and negative filters.

E. Bulk send/receive

Pull API is able to provide client with requested number of events on one call, however here server is at command at maximum limit of events it is willing to send. If we allow bulk transfer for push API, server may receive arbitrary number of events – even very large number – in one call. As server has to balance throughput and responsiveness, it also has to do some limiting. We will thus let server to present client with limit constants (for both directions) in initial handshake communication, and also in error message structures, should the client overflow these constants.

F. Authentication

To mitigate possibility of impersonation among clients on one machine, clients will have to supplement shared secret (instead of their publicly known name) during queries. As the connection is always encrypted over HTTPS and shared secret is distributed only once on client registration over secure channel, there is no need to complicate things with additional encryption or handshake scheme. However note that this mechanism is only for transition phase to specifically tailored certificates, which will contain client (not only machine) identifier directly.

Clients will also have to have server authority certificate (or chain) at their disposal to be able to verify server authenticity.

G. State

At the server, each event gets assigned integer serial number. These numbers are sequential, so we can keep track of the last event “id” each client have received and next time provide him only with yet unseen events.

Server will also keep state of the last downloaded event for each client, thus freeing clients from necessity to keep permanent state themselves – however clients are free to provide their own notion of state id for each query and saving it on their own, should the need arise.

H. Platform

Experience shows that Perl is not ideal choice anymore, however we would like to stay with flexibility, rapid prototyping and deployment speed of dynamic scripting language. We need solid library support for JSON and HTTPS on client side and support for high performance data based (non HTML) web application, along with good database support.

As Warden client library is only a communication channel, on which other third party data processing applications will be based, we also have to consider user base scope, libraries and frameworks support.

Our choice fell naturally to Python, based also on experiences on other projects. Python standard library already provides HTTP and HTTPS support, work with X509 certificates for authentication, WSGI support for connectors to powerful web server software, solid database support, and also handful of scientific frameworks (namely NumPy [8], SciPy [9] and Matplotlib [7]).

IV. HTTP API DESIGN

Leaving SOAP creates possibilities for arguments and results representation. We can identify two three classes of transferred data – structured event data (transferred both directions), error explanations (only from server to client) and query modification arguments (only from client to server).

Considering modification arguments, such as authentication tokens, filtering and first event ID, the well understood and widely used notion of URL parameters suits well – both sides know the type of the value, so we only need to transfer key/value pairs of strings. Repeated arguments can easily mimic multivalued/arrays.

For structured data in push direction POST data can be used and for pull direction we can send resulting data directly. However, we will have to settle for structure.

The examples are provided as calls to command line HTTP client utility *curl* [2], which also shows that by using this design we are able to access server methods even without client library, which is very useful for debugging, and can be also used as a base for very lightweight clients.

A. Error handling

If HTTPS call succeeds (200 OK), method returns JSON object containing requested data.

Should the call fail, server returns HTTP status code, together with JSON object, describing the errors (there may be multiple ones, especially when sending events). The keys of the object, which may be available, are:

- *method* - name of the called method
- *req_id* - unique identifier or the request (for troubleshooting, Warden administrator can also uniquely identify related log lines)
- *errors* - always present list of JSON objects, which contain:
 - *error* - HTTP status code
 - *events* – list of indices of events, affected by this particular error. If there is error object without *events* key, caller must consider all events affected
 - *message* - human readable error description

Other context dependent fields may appear, see particular method description.

Client errors (4xx) are considered permanent - client must not try to send same event again as it will get always rejected - client administrator will need to inspect logs and rectify the cause.

Server errors (5xx) may be considered by client as

temporary and client is advised to try again after reasonable recess.

B. Common arguments

- *secret* - shared secret, assigned to client during registration
- *client* - client name, optional, can be used to mimic Warden 2 authentication behaviour if explicitly allowed for this client by server administrator

C. *getEvents* method

Fetches events from server.

1) Arguments

- *count* - number of requested events
- *id* - starting serial number requested, id of all received events will be greater
- *cat, nocat* - selects only events with categories, which are/are not present in the event Category field (mutually exclusive)
- *group, nogroup* - selects only events originated/not originated from this realms and/or client names, as denoted in the event Node.Name field (mutually exclusive)
- *tag, notag* - selects only events with/without this client description tags, as denoted in the event Node.Type field (mutually exclusive)

2) Returns

- *lastid* - serial number of the last received event
- *events* - array of IDEA events

3) Example

```
$ curl \
  --key key.pem \
  --cert cert.pem \
  --cacert ca.pem \
  --request POST \
  \
  "https://warden.example.org/getEvents?\
secret=SeCrEt\
&count=1\
&nogroup=org.example\
&cat=Abusive.Spam\
&cat=Fraud.Phishing"

{"lastid": 581,
 "events": [{
  "Format": "IDEA0",
  "DetectTime": "2015-02-03T09:55:21.563638Z",
  "Target": [{"URL": ["http://example.com/"]}],
  "Category": ["Fraud.Phishing"],
  "Note": "Example event"
}]}
```

D. *sendEvents* method

Uploads events to server.

1) Arguments

- POST data - JSON array of Idea events

2) Returns

Object with number of saved messages in *saved* attribute.

3) Example:

```
$ eventId=$RANDOM$RANDOM$RANDOM$RANDOM$RANDOM
$ detecttime=$(\
date --rfc-3339=seconds|tr " " "T")
$ client="cz.example.warden.test"
$ printf '
[
{
  "Format": "IDEA0",
  "ID": "%s",
  "DetectTime": "%s",
  "Category": ["Test"],
  "Node": [{"Name": "%s"}]
}
]' $eventId $detecttime $client | \
curl \
  --key $keyfile \
  --cert $certfile \
  --cacert $cafile \
  --request POST \
  --data-binary "@-" \
  "https://warden.example.org/sendEvents?"\
  "client=$client&secret=SeCrEt"

{"saved":1}
```

4) Example with error:

```
$ curl \
  --key $keyfile \
  --cert $certfile \
  --cacert $cafile \
  --connect-timeout 3 \
  --request POST \
  --data-binary '[{"Format": "\
  "IDEA0", "ID": "ASDF", "Category": [], "\
  "DetectTime": "asdf"}]' \
  "https://warden.example.org/sendEvents?"\
  "client=cz.example.warden.test&secret=SeCrEt"

{"errors":
[
  {"message": "Validation error:
key \"DetectTime\", value \"asdf\", expected -
RFC3339 timestamp.",
  "events": [0],
  "error": 460
}
],
"method": "sendEvents",
"req_id": 3726454025
}
```

E. getInfo method

Provides client with basic server information.

1) Returns

- *version* - Warden server version string
- *description* - server greeting
- *send_events_limit* - sendEvents will be rejected if client sends more events in one call
- *get_events_limit* - getEvents will return at most that much events

2) Example

```
$ curl \
  --key key.pem \
  --cert cert.pem \
  --cacert ca.pem \
  --connect-timeout 3 \
  --request POST \
  "https://warden.example.org/getInfo?
secret=SeCrEt"

{"version": "3.0-beta1",
"send_events_limit": 500,
"get_events_limit": 1000,
"description": "Warden 3 server"}
```

V. DATABASE DESIGN

A. Essential queries

Lets take a look at the queries, which will create focus of the server work.

Each pull query is based on *id*, provided by client, signalling which events it has already received. Clients may also require to filter events according to category, detector tags and trailing part of detector name (“*realm*”).

Push queries are nothing special, they will just have to update all potential auxiliary structures accordingly – but the backend database engine must provide enough locking granularity to be able to cope with continuous stream of writes along with continuous stream of reads.

Each query has to be authenticated by client shared secret and/or client name, and we should be able to differentiate between clients, which are allowed to send, clients, which are allowed only to receive, and new, unverified clients, which are able to send only events marked with specific “Test” category.

B. Discussion

It is clear that we need the table of events and table of clients and positive relation between them.

Concerning events table, the only information we need to parse out from arriving JSON events are the filtering fields, so we do not need to try to represent the whole IDEA structure in database. However, this is in fact mainly Category array, which means we have one to many relation and we will have to split these into separate table. The same applies to the detector tags.

Both tags and categories are transferred as free text strings, which shows as a performance and space intensive way to represent them in relations. We have also tried conversion to hashed fixed length strings, which lessened impact, however we have still felt, that there is a margin. However database itself does not need to work with text identifiers directly in the queries, so we will create mapping of these finite sets to integer sequence and use these quite short integers in the database representation.

Interesting situation arises in connection with detector names – we have to be able to filter by prefix substring – “*org*”, “*org.example*”, “*org.example.honeypot*” can all be used as patterns. One possibility is to create auxiliary table with all possible prefixes, however that shows very bad performance in negative queries, where database is forced to generate large list of all non matching prefixes on which it

consequently filters event data. However database indices are indeed prefix based, so correctly used anchored *LIKE* operator may be enough.

Next complication we have to solve is saving last pull ids of accessing clients – straightforward solution would be to store it in the table of clients. However, table of clients is mostly immutable, and from administration point of view it would be wise to leave access to it only to human operator, avoiding frequent changes by server itself, we will thus split this information into *last_events* table in the form of client identifier, last event id and login timestamp.

Concerning database engine itself – we prefer raw performance over capabilities – we can miss a few security events in case of outage if server is able to withstand large number of concurrently accessing client connections. Our first choice points to MySQL, whose InnoDB engine supports reasonable number of database capabilities and fine grained line based locking (necessary for concurrent read/write access) together with decent performance.

C. Final schema

```
CREATE TABLE events (
  id int(11) NOT NULL AUTO_INCREMENT,
  received timestamp NOT NULL,
  client_id int(11) NOT NULL,
  `data` longtext NOT NULL,
  valid tinyint(1) NOT NULL DEFAULT 1,
  PRIMARY KEY (id),
  KEY id (id,client_id)
);

CREATE TABLE event_category_mapping (
  event_id int(11) NOT NULL,
  category_id int(11) NOT NULL,
  KEY event_id_2 (event_id,category_id)
);

CREATE TABLE event_tag_mapping (
  event_id int(11) NOT NULL,
  tag_id int(11) NOT NULL,
  KEY event_id_2 (event_id,tag_id)
);

CREATE TABLE clients (
  id int(11) NOT NULL AUTO_INCREMENT,
  registered timestamp NOT NULL,
  requestor varchar(256) NOT NULL,
  hostname varchar(256) NOT NULL,
  note text NULL,
  valid tinyint(1) NOT NULL DEFAULT 1,
  name varchar(64) NOT NULL,
  secret varchar(16) NULL,
  `read` tinyint(1) NOT NULL DEFAULT 1,
  `write` tinyint(1) NOT NULL DEFAULT 0,
  test int(11) NOT NULL DEFAULT 0,
  PRIMARY KEY (id)
);

CREATE TABLE last_events (
  id int(11) NOT NULL AUTO_INCREMENT,
  client_id int(11) NOT NULL,
  event_id int(11) NOT NULL,
  `timestamp` timestamp NOT NULL,
  PRIMARY KEY (id),
  KEY client_id (client_id,event_id)
);
```

VI. PERFORMANCE

Nowadays there is 36 millions of events in the Warden database of average event size 786.3 B. Clients produce on average 7 events per connection, however bunches of 1000 events (selected as reasonable maximum) are not rare. However median is 2 events per connection, so we will have to make sure final Warden 3 performs well both on many small accesses and on bulk uploads.

Indicative testing of prototype, based on this design, shows that practical limit may be somewhere about 40 000 incoming events per second on single event per connections, however throughput seems to be able to raise at least up to 110 000 events per seconds when clients use 100 events per connection – this indicative test was made using up to 80 simultaneously accessing clients and shows that outlined design is worth pursuing.

VII. CONCLUSION

Warden 3 is complete redesign, based on the identified shortcomings emerged during several years of Warden 2.X operation. Which is not to lessen merit of Warden 2, it is necessary to note that without it, Warden 3 wouldn't most probably exist. New Warden uses flexible and descriptive event format, based on JSON. Warden 3 protocol is based on plain HTTPS queries with help of JSON (Warden 2 SOAP is heavyweight, outdated and draws in many dependencies). Clients can be multilanguage, unlike SOAP/HTTPS, plain HTTPS and JSON is mature in many mainstream programming languages. Server is written in Python - mature language with consistent and coherent libraries and many skilled developers, and uses MySQL as efficient data storage.

VIII. FUTURE WORK

Some of the Warden 2 clients were already converted to prototype Warden 3 instance with superb results, however turning the prototype into full production state along with subsequent transferring of existing clients will still need a lot of work.

However, family of tools, based on Warden client library, will be able to emerge, namely connectors for Kippo and Dionaea honeypots, connectors to storage and data mining tools and many others – able to work with much wider pool of information, accessible in Warden's new extensive security event format.

IX. ACKNOWLEDGMENT

The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum (<http://www.metacentrum.cz/en/>), provided under aforementioned programme, is highly appreciated.

REFERENCES

- [1] V. Bartoš, M. Žádník, An Analysis of Correlations of Intrusion Alerts in an NREN. *19th International Workshop on Computer-Aided Modeling Analysis and Design of Communication Links and Networks (CAMAD)*. IEEE, Athens, December 2014. ISBN: 978-1-4799-7134-3.
- [2] *cURL*. Cited 29 May 2015. Available: <http://curl.haxx.se/>
- [3] P. Kácha, IDEA: Designing the Data Model for Security Event Exchange, *17th International Conference on Computers: Recent Advances in Computer Science*, Rhodes, 16 July 2013, ISBN: 978-960-474-311-7, ISSN: 1790-5109. Available: <http://www.wseas.us/e-library/conferences/2013/Rhodes/COMPUTE/COMPUTE-36.pdf>.
- [4] P. Kácha, IDEA: Security Event Taxonomy Mapping, *18th International Conference on Circuits, Systems, Communications and Computers*, Santorini, 17 July 2014, ISBN: 978-1-61804-236-1, ISSN: 1790-5109. Available: <http://www.europment.org/library/2014/santorini/bypaper/COMPUTERS/COMPUTERS1-19.pdf>
- [5] P. Kácha, *Incident Classification Comparison (with eCSIRT.net mkII as main reference)* [online], CESNET, 10 January 2014. Available: https://csirt.cesnet.cz/media/en/idea/incident_classification_comparison.ods
- [6] J. Mach, *Expert system Mentat*, CESNET 9 December 2013. Available: <http://www.cesnet.cz/wp-content/uploads/2013/12/mentat-paper.pdf>
- [7] *matplotlib*. Cited 29 May 2015. Available: <http://matplotlib.org/>
- [8] *NumPy*. Cited 29 May 2015. Available: <http://www.numpy.org/>
- [9] *SciPy*. Cited 29 May 2015. Available: <http://www.scipy.org/>
- [10] D. Stikvoort, *Incident Classification* [online]. 23 May 2013. Available: <http://www.terena.org/activities/tf-csirt/meeting39/20130523-DV1.pdf>
- [11] *Warden*. CESNET. Copyright 2010-2013. Last updated 17 April 2013. Available: <https://wardenw.cesnet.cz/en/index>

Ransomware

Jan Kolouch, Andrea Kropáčová

Abstract—A combined attack, frequently referred to as ransomware, has recently become one of most common cyber attacks. This attack combines social engineering, malware attack and a criminal offence in the form of blackmail. The offender’s aim is to gain a pre-defined “ransom” from as many end users as possible. Ransomware is a part of “malware economy” in which a number of similar attacks is launched with the aim of gaining profit from the victims. The article below defines ransomware and discusses the international legal liability for such types of attack.

Keywords—Ransomware, cybercrime, Convention of Cybercrime, criminal liability, reverse analysis, social engineering.

I. INTRODUCTION

At present, combined attacks, frequently referred to as *ransomware*, have recently expanded dynamically. Such attacks cannot be easily defined as for instance a DoS or DDoS attack [1], [2], probe, scanning, hacking, or similar attacks. Ransomware attack combines elements of social engineering (luring the victim into downloading the infected file or visiting malicious websites), malware attack (infecting victim’s personal computer and taking control over it) and criminal offence in the form of blackmail, the aim of which is generally to gain a pre-defined “ransom” from as many users as possible.

There is no doubt that ransomware can be classified as *internet organised crime*, or an element of the “malware economy”. Ransomware in fact consists of a series of very similar attacks, the aim of which is to profit on as many victims as possible, irrespective of victim’s location in the digital world, age, social background, education, etc.

II. CYBER ATTACK

Prosiše and Mandiva define a “**computer security incident**” (that can be perceived as a cyber attack or cyber crime) as an unlawful, illegal, unauthorised, unacceptable action that concerns a computer system or a computer network. Such action can take the form of for instance personal data theft, spam or other intrusion, misappropriation, proliferation or possession of child pornography and others [3].

A cyberattack is deliberate exploitation of computer systems, technology-dependent enterprises and networks.

= This work has been supported by the CESNET, a. l. e., <http://www.cesnet.cz>, operator of the Czech national research and education network referred to as CESNET2 within its “Large Infrastructure” (LM2010005) research programme of the Ministry of Education, Youth and Sports of the Czech Republic, running within 2010-2015 timeframe, .

Jan Kolouch works in CESNET, a. l. e., Zikova 4, Prague 6, Czech Republic, (email: kolouch@cesnet.cz).

Andrea Kropáčová works in CESNET, a. l. e., Zikova 4, Prague 6, Czech Republic, (email: andrea@cesnet.cz).

Cyberattacks use malicious code to alter computer code, logic or data, resulting in disruptive consequences that can compromise data and lead to cybercrimes, such as information and identity theft [4].

Consequently, a **cyber attack**¹ can also be defined as **any illegal action by the offender in the cyberspace targeted against the interests of another person**. Such action does not always constitute a criminal offence; the key is that it hinders everyday life of the injured. A cyber attack can be either completed or it can be in preparation or attempted only.

Cyber criminality takes the form of cyber attacks. Cyber criminality is the crime in which IT technology is [4]:

- a) *used as the tool to commit a criminal offence,*
- b) *The target of offender’s attack. The attack constitutes a criminal offence provided IT technology is used or misused in the information, system, programming or communication environment.*

Certain illegal action in the cyberspace or action related to cybercrime can be classified according to relevant provisions of country’s regulations. There is, however, certain action that is difficult or impossible to classify as a criminal offence.

III. RANSOMWARE

Many authors perceive **ransomware** as a type of malware. *Ransomware is a type of malware that prevents or limits users from accessing their system. This type of malware forces its victims to pay the ransom through certain online payment methods in order to grant access to their systems, or to get their data back. Some ransomware encrypts files (called Cryptolocker). Other ransomware use TOR to hide C&C communications (called CTB Locker) [5].*

Ransomware deploy malicious software which can disable the functionality of your computer [6].

Ransomware is malware for data kidnapping, an exploit in which the offender encrypts the victim’s data and demands payment for the decryption key [7].

Ransomware that locks a computer and uses law enforcement imagery to intimidate victims has spread from Eastern Europe to Western Europe, the United States, and Canada over the past year². The scam has been copied and professionalized from initial early attacks, with established online criminal gangs now branching out into the scheme. Each gang has separately developed, or bought, their own

¹ A cyber attack does not correspond with the term ‘**security incident**’. A security incident is the violation of IS/IT security and the rules designed to protect it (security policy).

² Read Symantec report from year 2012.

different version of the ransomware. This malware is highly profitable, with as many as 2.9 percent of compromised users paying out. An investigation into one of the smaller players in this scam identified 68,000 compromised computers in just one month, which could have resulted in victims being defrauded of up to \$400,000 USD. A larger gang, using malware called Reveton (aka Trojan.Ransomlock.G), was detected attempting to infect 500,000 computers over a period of 18 days. Given the number of different gangs operating ransomware scams, a conservative estimate is that over \$5 million dollars a year is being extorted from victims. The real number is, however, likely much higher [8].

The above quote merely demonstrates the success and mass proliferation of cyber attacks that can be classified as ransomware. **Production and distribution of ransomware is directly linked to activities referred to as Crime-as-a-Service.** Over the past few decades the digital underground has evolved and matured from a few small groups hacking and phishing for fun and prestige, to a thriving criminal industry that costs global economies an estimated USD 300+ billion per year [9].

Crime-as-a-Service can be described as a business model (toolkit) which may include malicious software, supporting infrastructure, stolen personal and financial data and the means to monetize their criminal gains. With every aspect of this toolkit available to purchase or hire as a service, it is relatively easy for cybercrime initiators – lacking experience and technical skills – to launch cyber attacks not only of a scale highly disproportionate to their ability but for a price similarly disproportionate to the potential damage [10], [11].

Both from the professional and legal point of view, the above described actions represent a simplification of the complex process of the cyber attacks qualified as ransomware. This is why this article first deals with one of the best known ransomware – “**Police ransomware**”. Next, the punishment of the offender of the action with elements of ransomware within the framework of the international law will be discussed.

IV. “POLICE RANSOMWARE”

Since 2011, ransomware attacks on end user computers have taken place in almost every EU Member State. The aim of these attacks is to gain financial profit. The essence of the attacks is that the victim’s computer is infiltrated with malware. Thus, the infected computer becomes a part of a botnet, through which the “blackmailing virus” spreads further. Subsequently, malware blocks access to the account of the OS Windows user while notifying him that the computer was blocked by the relevant country’s police (see Figure 1).

In this case, the offenders seek to abuse people’s trustfulness and “the appearance” of an official authority’s communications to gain money from the users.

V. LEGAL ASPECTS OF RANSOMWARE

The core international document defining cyber offences is the Council of Europe’s **Convention No. 185 on**

Cybercrime from 23 November 2001 [12].

The Convention on Cybercrime constitutes the first international agreement in respect of criminal offences committed by means of information technologies (the misuse of the Internet and other computer network in particular) such as copyright infringement, computer fraud, proliferation of child pornography and other types of attacks against information and computer security. The aim of the Convention is to streamline the approach of Convention’s signatories towards sanctioning the most serious types of cyber attacks.

From the legal point of view, a ransomware attack generally consists of the following actions:

1. sending an infected file, or a link to an infected website;
2. entering the malicious code in the computer;
3. data encryption in order to restrict the rightful user’s access to the data or the system;
4. a request to pay a certain amount to unblock the computer or the data.

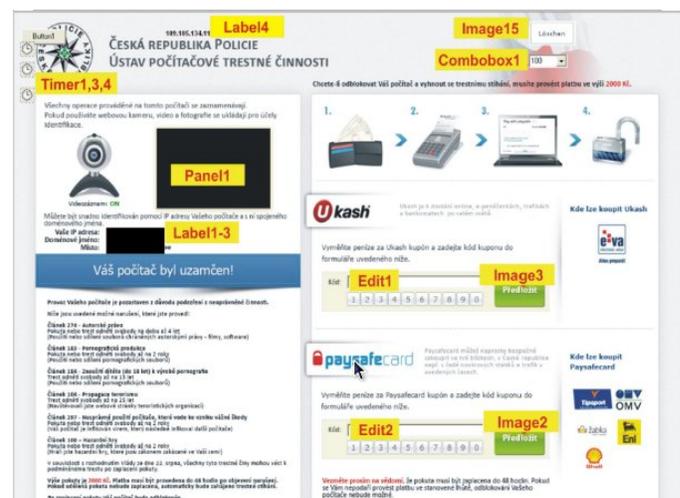


Figure 1 – An fake www page of Czech Police.

Ad 1. Sending an infected file, or a link to an infected website

The attack itself is de facto based on a phishing attack. Most frequently, the victim is sent an email which on the first sight does not raise any suspicion that it could contain a fake message. In general, such emails contain a link which the user is prompted to follow. Once the user has clicked on the attached link, he is directed to a website, the layout and functions of which do not differ from the authentic website. In reality, the user is redirected to a fake website which imitates the original more or less faithfully. Phishing collects data entered on the fake websites and sends them automatically to the offender. This way, the offender can obtain identification data of internet banking system’s users or access to individual bank accounts of the users of the infected systems. The offender can also obtain identification numbers and other data relating to the payment cards which enable him to subsequently make payments through the Internet etc.

During the ransomware attack, the user is not requested to make a payment or signed and administer secured accounts, etc. The aim of the offender is to persuade the user to visit the website on which the fake code is located. Once the user accesses this website, an attack on his computer is launched. This attack enables the offender to take over the control over the infected computer, to install malware to be able to control it distantly. Very often, the users' personal computers are attacked, compromised (e.g. by malware) and become a part of a *botnet*. The botnets enable generating massive DoS/DDoS attacks, or hide the activities of the offender and his identity in executing more sophisticated and precisely targeted attacks with a severe impact, see also [13].

Enclosing the fake code directly in the e-mail is another way to infect victim's computer. In such cases, the offender frequently misuses the default OS settings (mostly Windows based OS) enabling to hide the known file types. The victim for instance believes that he is opening a file in Word while his computer is being infected by means of malware (see Figure 2).

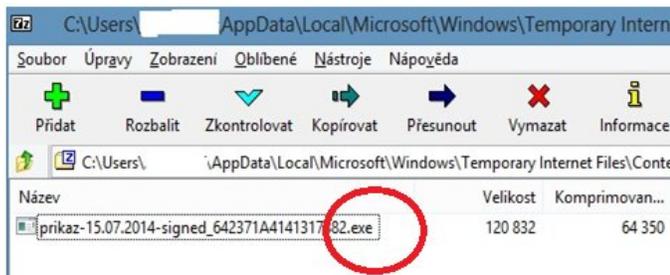


Figure 2 – An executable programme hidden as an order from a public authority

From the legal point of view, the **Convention on Cybercrime** classifies the action by the offender, i.e. sending of the file through which the offender may gain control over somebody else's computer, or re-directing to the website containing malware, as an *attempt* or *aiding or abetting* to criminal offences. In this case, the action most likely constitutes an attempt to commit a criminal offence as defined in Articles 4 through 6 of the Convention on Cybercrime. For future reference, the above mentioned articles of the Convention on Cybercrime are described below in detail:

Article 4 of the Convention – Data interference

1 Each Party shall adopt such legislative and other measures as may be necessary to establish as criminal offences under its domestic law, when committed intentionally, the damaging, deletion, deterioration, alteration or suppression of computer data without right.

2 A Party may reserve the right to require that the conduct described in paragraph 1 result in serious harm.

In conjunction with the relevant provisions of national criminal law, this article provides, for sanctioning actions consisting of **intentional installation of malware into a computer system without the consent of the system's rightful user.**

Article 5 of the Convention – System interference

Each Party shall adopt such legislative and other measures as may be necessary to establish as criminal offences under its domestic law, when committed intentionally, the serious hindering without right of the functioning of a computer system by inputting, transmitting, damaging, deleting, deteriorating, altering or suppressing computer data.

While Article 4 of the Convention defines the merits of a criminal offence against data in a computer system, i.e. the interference with the data does not necessary cause damage the computer system (e.g. changing data in a database), this Articles protects the functioning of a computer system as a whole, and the actions described in Article 4 here hinder the functioning of the computer system affected.

Article 6 of the Convention – Misuse of devices

1 Each Party shall adopt such legislative and other measures as may be necessary to establish as criminal offences under its domestic law, when committed intentionally and without right:

a the production, sale, procurement for use, import, distribution or otherwise making available of:

i a device, including a computer program, designed or adapted primarily for the purpose of committing any of the offences established in accordance with Articles 2 through 5;

ii a computer password, access code, or similar data by which the whole or any part of a computer system is capable of being accessed,

with intent that it be used for the purpose of committing any of the offences established in Articles 2 through 5; and

b the possession of an item referred to in paragraphs a.i or ii above, with intent that it be used for the purpose of committing any of the offences established in Articles 2 through 5. A Party may require by law that a number of such items be possessed before criminal liability attaches.

In accordance with the above provision, **all offenders who proliferate**, sell, procure for themselves or other, import, distribute or otherwise make available for instance malware (programme such as computer worms, Trojan horses, key loggers, etc) should be sanctioned. Article 6 of the Convention basically reclassifies the possession and handling of such programmes from an attempt to a completed criminal offence. This provision, however, does not provide for the sanctioning of the production of such computer programmes, unless the specific offender's intent (to commit any of the above listed criminal offences under the provisions of Articles 2 through 5 of the Convention) has been proved.

This provision should not be interpreted as if extending the criminal liability to each single disposal with the listed software. **For the criminal liability to arise, such tools should be possessed with the intent to commit any of the above listed criminal offences as defined in Articles 2 through 5 of the Convention.** Similarly, these provisions do not cover situations in which the protection of a computer system against malicious software is tested through a deliberate exposure of the computer system to such threats by the authors of the security measures.

Ad. 2 Entering the malicious code in the computer

From the legal point of view, the action by the offender

consisting of the malware installation (without the consent of the rightful user) into the compromised device constitutes a completed criminal offence as defined in Articles 2, 4 and 5 of the Convention on Cybercrime. Article 2 of the Convention defines ‘**Illegal access**’ as committed by a person through gaining an unauthorised access to a computer system or its part.

It follows from the wording of the Article that the codification of the circumstances listed in domestic law is optional. Unless the circumstances are enacted as the indispensable condition for offender’s criminal liability for the commitment of such criminal offence, a mere gaining of unauthorised access to a computer system or its part should be criminally punishable. Such regulation would enable to criminally punish for instance hacker attacks consisting of merely gaining the access to a computer system, even where the hacker’s action caused no harm or where the hacker had not manipulated with data and information he obtained or got acquainted with during the attack.

Ad. 3. Setting a password on (hindering access to) rightful user’s data

The action of the offender consisting in setting a password on (hindering access to) data to prevent the rightful user from accessing them can be classified as action described in Articles 4 and 5 of the Convention on Cybercrime. The merits of such action are deemed fulfilled once the user is hindered from a free use of own computer and data stored in it. The blocking of the computer itself is directly linked to the action described in paragraph below.

Ad. 4. Request for the payment of ransom to unblock the computer or data

The primary goal of the entire attack, shortly described as ransomware, is to gain financial profit for “unblocking the computer or data” for the compromised user. In general, the user is requested to send a certain financial amount, either through any of the payment portals (e.g. Ukash, paysafecard, MoneyPak, etc.) or through Bitcoin. Once the sum has been paid, a key (chain) renewing the access to the computer system or user data, is provided. However, malware level remains installed in the computer. Thus, the offender may repeat his request (blackmailing).

From the legal point of view, the above described action can be classified as blackmail, since the offender forces another person to act, neglect or sustain something. We presume that such action could be also subsumed under **Article 8 of the Convention on Cybercrime - Computer-related fraud.**

Each Party shall adopt such legislative and other measures as may be necessary to establish as criminal offences under its domestic law, when committed intentionally and without right, the causing of a loss of property to another person by:

- a) *any input, alteration, deletion or suppression of computer data,*
- b) *any interference with the functionality of a computer system, with fraudulent or dishonest intent of procuring, without right, an economic benefit for oneself or for another person.*

This provision defines a specific type of fraud, or more precisely a fraud committed in a specific manner – by interfering with computer data or computer system’s functions.

The above described action, which according to this Article should be criminally punishable, occurs most frequently in conjunction with other action that the Convention aims to mitigate. For instance, the offender first obtains the programme that enables him to interfere with a computer system without authorisation (Article 6). Next, he uses the programme obtained to execute the attack by simulating the person’s authorisation to dispose with a bank account (Articles 4 and 7). Finally, he may give instructions to transfer money to his benefit or to the benefit of a third party (Article 8).

VI. CONCLUSION

Based on the above analysis, we believe that technical and legal professionals should cooperate more closely in the fight against cyber attacks or cybercrime. This would enable to revise the legal regulations of individual member states to sanction unwanted Internet action and to enable CERT/CSIRT teams³ and law enforcement agencies in particular to fully exploit the means and the limits of the law to repress such illegal action.

REFERENCES

- [1] Andrea Kropáčová, (D)DoS attacks targeted web servers operated in Czech Republic, 17th International Conference on Computers: Recent Advances in Computer Science, Rhodos, 16 July 2013, ISBN: 978-960-474-311-7, ISSN: 1790-5109
- [2] Jan Kolouch, “Criminal liability for DoS and DDoS attacks”, 17th International Conference on Computers: Recent Advances in Computer Science, Rhodos, 16 July 2013, ISBN: 978-960-474-311-7, ISSN: 1790-5109
- [3] PROSISE, Chris, MANDIVA, Kevin. *Incident response & Computer forensic, second edition*. Emeryville : McGraw-Hill Companies, 2003. p. 13. Compare also CASEY, Eoghan. *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet, Second Edition*. London : Academic Press, 2004. p. 9 and further.
- [4] KOLOUCH, Jan and Petr VOLEVECKÝ. *Criminal protection against cyber crime*. Prague: The Police Academy of the Czech Republic in Prague, 2013. ISBN 978-80-7251-402-1. p. 12.
- [5] <http://www.trendmicro.com/vinfo/us/security/definition/ransomware>
- [6] <http://www.trendmicro.com/vinfo/us/security/definition/ransomware>
- [7] <http://us.norton.com/ransomware>
- [8] <http://whatis.techtarget.com/definition/ransomware-cryptovirus-cryptotrojan-or-cryptoworm>
- [9] http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/ransomware-a-growing-menace.pdf
- [10] <http://www.mcafee.com/nl/resources/reports/rp-economic-impact-cybercrime.pdf>
- [11] https://www.europol.europa.eu/content/internet-organised-crime-threat-assessment-iocta_page_20
- [12] http://www.strategicstudiesinstitute.army.mil/pdf/files/PUB1145.pdf_page_15
- [13] Convention on Cybercrime CETS No. 185:
<http://conventions.coe.int/Treaty/Commun/QueVoulezVous.asp?CL=ENG&NT=185>
- [14] Daniel Plohmann, Elmar Gerhards-Padilla, Felix Leder. *Botnets: Measurement, Detection, Disinfection and Defence*. [online]. [28. 3. 2014]. Available at: <http://www.enisa.europa.eu/activities/Resilience-and-CIIP/critical-applications/botnets/botnets-measurement-detection-disinfection-and-defence>

³ CERT = Computer Emergency Response Team, CSIRT = Computer Security Incident Response Team.

Face Depth Estimation Using Differential Evolution and Iterative Soft Thresholding Algorithm

K. Punnam Chandar
 Dept. of E.C.E
 Kakatiya University
 Warangal, INDIA
 k_punnam@yahoo.co.in

T. Satya Savithri
 Dept. of E.C.E
 Jawaharlal Nehru Technological University
 Hyderabad, INDIA
 tirumalasadtya@gmail.com

Abstract—In this paper we propose a novel face features depth estimation algorithm based on similarity transform. The formulated problem is consisting of two stages. In the first stage the head pose is estimated using Differential Evolution optimization (DE). In the second stage the results of the first stage are used to estimate the depth values of important features using Iterative Soft Thresholding algorithm (ISTA). Experimentation is carried on 3D Bosphorus Database as it provides the 2D coordinates and corresponding 3D coordinates and fair comparison of the depth values estimation can be made. Further to show the Depth Estimation efficacy of the proposed algorithm Similarity Metric, Pearson Linear Correlation coefficient is computed and tabulated. All the simulations were carried out in MATLAB and the results are satisfactory.

Keywords—CANDIDE model; Differential Evolution; Iterative Soft Thresholding Algorithm; Structure-from-motion; 3D face Reconstruction;

I. INTRODUCTION

2D-face recognition systems have been developed with more than three decades of research but exhibit well known deficiencies. Only under controlled conditions the 2D-face recognition systems achieve reasonable performance level [1-2]. 2D wide angle Surveillance cameras are deployed to enhance security requirements in small-scale standalone applications to large scale networked Closed Circuit Televisions in law enforcement for public streets. The use of wide angle cameras maximizes the viewing area at the cost of decrease in performance as the controlled conditions are hard to meet. In real time with the query image having arbitrary pose with very small face region when the person is far away from the camera the performance of 2D face recognition systems suffers dramatically.

One way to improve the Face recognition performance under arbitrary pose is to use multiple training images under different poses. In real time multiple face images of the subject under consideration may not be immediately available and use of multiple face images will increase the storage and computation time required for further useful analysis.

A potential method to improve the recognition performance in low quality input video data is to employ super resolution algorithms prior to face recognition. For these methods to give satisfactory performance, the low

quality images should comply data constraints like visual quality and face similarity and will be computationally burden. Therefore 3D face data acquisition and shape representation research is gaining momentum. Gradually the 3D face models have been developed, to use in face recognition, face tracking and face animation, etc. A comprehensive review is presented in [3] for 3D and multi-modal 3D+2D face recognition.

Presently, there are two main stream approaches usually adopted to create human facial 3D models. One of the approach is to use specialized 3D scanners to capture texture in addition to depth i.e., 3D shape. The high cost and speed limitations of 3D scanning devices are the obvious short comings to acquiring sufficient and useful data for processing and further useful analysis. The second approach is to develop algorithms to reconstruct 3D face models from 2D images, such as video sequences [5] or multi-view photographs [6]. The second way of 3D reconstruction mainly depends on the development of novel algorithms utilizing the existing 2D surveillance systems. When 3D face models are constructed from low quality data, the critical issues to be considered are utilization of available prior information and the design of efficient reconstruction algorithm. An efficient 3D reconstruction algorithm can greatly enhance the capabilities of existing 2D face recognition systems.

During the past decade many 3D reconstruction algorithms have been developed and can be classified in to three groups, shape-from-shading (SFS) [7, 8, 9, 15], the 3D Morphable model [10, 11, 12], and structure from motion [6, 13, 14]. We confine to presenting the details of the SFM as this form the basis of our work.

Structure from motion (SFM) is a popular approach to recover the 3D shape of an object when multiple frames of an image sequences are available. Given a set of observations of 2D feature points, SFM can estimate the 3D structure of the feature points by inverting the effect of the projection process. As from the literature two well-known projection models are available, the perspective and the orthographic. Perspective projection is a realistic model of the imaging process, whereas the orthographic projection yields easy-to-solve models that are applicable in some simple cases. Orthographic projection is used primarily because it gives rise to mathematically tractable equations.

Much research has been conducted on determining the motion and structure of moving objects under orthographic projection. Ullman [16] proved that four point correspondences over three views yield a unique solution to motion and structure. Tomasi [13] proved the rank-3 theorem, i.e., the rank of the observation matrix is 3 under an orthographic projection, and proposed a robust factorization algorithm to factor the observation matrix into a shape matrix and a camera motion matrix using the singular value decomposition (SVD) technique. Xirouhakis and Delopoulos [17] extracted the motion and shape parameters of a rigid 3D object by computing the rotation matrices via the eigenvalues and eigenvectors of appropriate defined 2x2 matrices, where the eigenvalues are the expression of four motion vectors in two successive transitions. Bregler [18] assumed a 3D object to be non-rigid, and the observed shapes are represented as a linear combination of a few basis shapes. Further, in [19], a Gaussian prior is assumed for the shape coefficients, and the optimization is solved using the expectation-maximization [EM] algorithm. In [6] a similarity-transform-based (SM) method is proposed to derive the 3D structure of a human face from multi-view photographs. The SM algorithm does not require any prior knowledge of camera calibration, and has no limitation on the possible poses or the scale of the face images. In addition SM method has been verified that it can be extended to face recognition to alleviate the effect of pose variations. Unfortunately, the genetic algorithm (GA) is used in SM algorithm to estimate the depth and is usually encounters a heavy computational burden. Moreover, how to design a reasonable chromosome, how to make a feasible gene operation scheme and how to adjust the parameters remain difficult problems.

In this paper we propose optimization algorithm utilizing Differential Evolution (DE) [22] and Iterative Soft Thresholding algorithm (ISTA) [24, 28] to estimate the head pose and depth values of facial feature points i.e. a sparse 3D face representation from a single 2D image. In this approach, DE is used to estimate the head pose and followed by ISTA to estimate the depths of important face features shown in Fig.1. Additionally Candide model [20] widely used in face related applications [6, 21] is used as initial face structure to improve the accuracy of the algorithm.



Fig.1 Feature point positions in a frontal-view face image.

The proposed algorithm is tested on 3D Bosphorus Database, where by a fair comparison of the depth value estimation can be made. The remainder of this paper is organized as follows: In section II, we present our methodology; Experimental results are given in Section III, and conclusion in Section IV.

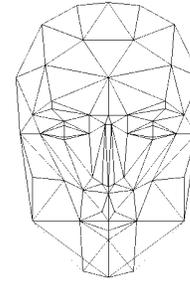


Fig 2 Candide-3 3D Face Model.

II. METHODOLOGY

Our algorithm requires only one frontal and one non-frontal view face image to estimate the 3D face structure. We assume that n shape feature points represented by

the $\sum_{i=1}^n (x_i, y_i)$ coordinates are marked accurately. (M_{xi}, M_{yi}, M_{zi}) represents the i^{th} feature point of a frontal-view 3D face model M. M_{xi} and M_{yi} are measured from the image being adapted, while M_{zi} is initially set at the default values of the CANDIDE model. (q_{xi}, q_{yi}) is a i^{th} feature point of a non-frontal-view 2D face q . The rotation matrix R for q is given as follows:

$$R_i = \begin{bmatrix} \cos\phi_i & \sin\phi_i & 0 \\ -\sin\phi_i & \cos\phi_i & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \cos\Psi_i & 0 & -\sin\Psi_i \\ 0 & 1 & 0 \\ \sin\Psi_i & 0 & \cos\Psi_i \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_i & \sin\theta_i \\ 0 & -\sin\theta_i & \cos\theta_i \end{bmatrix} = \begin{bmatrix} r_{i1} & r_{i2} & r_{i3} \\ r_{i4} & r_{i5} & r_{i6} \\ r_{i7} & r_{i8} & r_{i9} \end{bmatrix} \tag{1}$$

where the pose parameters Φ , Ψ , and θ are the rotation angles around the x, y, and z axes, respectively. Then the rotation and translation process for projecting the frontal-view face image to nonfrontal-view face image can be given by

$$\begin{pmatrix} q_{x_i} \\ q_{y_i} \end{pmatrix} = s \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \end{pmatrix} \begin{pmatrix} M_{x_i} \\ M_{y_i} \\ M_{z_i} \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix} \tag{2}$$

$i=1, \dots, p$

where s is the scale factor and (t_x, t_y) are the translations along x and y axes. The eq. (2) is Similarity Transform 3D and can be written in matrix form as follows:

$$q = sR_{2 \times 3}M + T \tag{3}$$

where q is a $2 \times p$ matrix such that each column represents the (x, y) coordinates $(q_{xi}, q_{yi})^T$ of one feature point, M is a matrix such that each column represents (x, y, z) the coordinates $(M_{xi}, M_{yi}, M_{zi})^T$ of one feature point, and t is a

2 x p matrix such that all columns are $(t_x, t_y)^T$. In terms of the shape-alignment approach in [6], the translation term \mathbf{t} can be eliminated if both \mathbf{q} and \mathbf{M} are centered at the origin, i.e.

$$\mathbf{q} \leftarrow \mathbf{q} - \bar{\mathbf{q}} \quad (4)$$

$$\mathbf{M} \leftarrow \mathbf{M} - \bar{\mathbf{M}} \quad (5)$$

$$\mathbf{q} = \mathbf{sR}_{2 \times 3} \mathbf{M} \quad (6)$$

Where $\bar{\mathbf{q}}$ is a 2 x p matrix such that each column is

$$\left(\frac{1}{p}\right) \sum_{i=1}^p (\mathbf{q}_{xi}, \mathbf{q}_{yi})^T \text{ and } \bar{\mathbf{M}} \text{ is a } 3 \times p \text{ matrix such that}$$

$$\text{each column is } \left(\frac{1}{p}\right) \sum_{i=1}^p (\mathbf{M}_{xi}, \mathbf{M}_{yi}, \mathbf{M}_{zi})^T.$$

Denote

$$\mathbf{A} = \mathbf{s} \cdot \mathbf{R}_{2 \times 3} \quad (7)$$

Equation (6) can then be rewritten as

$$\mathbf{q} = \mathbf{A} \cdot \mathbf{M} \quad (8)$$

If the pose of the face model and the depths of the feature points fit the non-frontal-view face image, the following similarity distance will be a minimum

$$\mathbf{d} = \|\mathbf{q} - \mathbf{A} \cdot \mathbf{M}\|_2^2 \quad (9)$$

A. Differential Evolution Optimization

Differential Evolution algorithm is a recently proposed real parameter optimization algorithm [22, 23] and its functioning is similar to the Genetic Algorithm and works through a simple cycle of stages shown in Fig. 1.

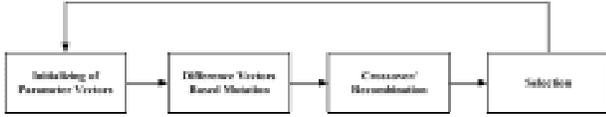


Fig. 3 Differential Evolution Algorithm Stages.

Differential Evolution like the method of Genetic Algorithm generates the initial poses that are randomly generated and evenly distributed to form the initial population. The fitness of each candidate in the population is measured based on eq. (9). To evolve new individuals that will be part of the next generation are created by combining members of the current population. Every individual acts as a parent and is associated to a donor vector. In the basic version of DE, the donor vector \mathbf{D}_i for the i^{th} parent (\mathbf{X}_i) is generated, by combining three random and distinct population members (\mathbf{X}_a), (\mathbf{X}_b) and (\mathbf{X}_c) as follows:

$$\forall i \in n : \mathbf{D}_i = \mathbf{X}_a + F(\mathbf{X}_b - \mathbf{X}_c) \quad (10)$$

where i, a, b, c are distinct

Where (\mathbf{X}_b) and (\mathbf{X}_c) are randomly chosen and (\mathbf{X}_a) is chosen either randomly or as one of the best

members of the population, F (Scale Factor) is a real -valued parameter that strongly influences DE's performance and typically lies in the interval [0.4, 1]. Other mutation strategies have been applied to DE, experimenting with different base vectors and different number of vectors for perturbation.

After mutation a trial vector $\mathbf{T}_{i,j}$ is generated by choosing between the donor vector and the previous generation for each element (j) according to the crossover rate CR which lies in the interval [0,1], for each element in the vector we choose either the corresponding element from the previous generation vector or from the donor vector such that

$$\forall i, j : \text{if } (\text{random} < CR \parallel j = j_{rand}) \text{ then} \quad (11)$$

$$\mathbf{T}_{i,j} = \mathbf{D}_{i,j} \text{ otherwise } \mathbf{T}_{i,j} = \mathbf{X}_{i,j}$$

Where j_{rand} is randomly chosen for each iteration through i and ensures that no \mathbf{T}_i is exactly the same as the corresponding \mathbf{X}_i . Then the trial vectors fitness is evaluated, and for each member of the new generation, \mathbf{X}_i , we choose the better performing of the previous generation, \mathbf{X}_i , or the trial vector, \mathbf{T}_i .

B. Iterative Soft Thresholding Algorithm

Differential Evolution optimization is used to estimate the rotation matrix \mathbf{R} for \mathbf{q} with initial depth values of candidate model in the first iteration, the estimated initial face structure is an approximation only. Therefore, to accurately estimate the depth values in \mathbf{M} , Iterative Soft Thresholding Algorithm is employed.

The orthographic projection model from 3D to 2D is

$$\mathbf{q} = \mathbf{A} \cdot \mathbf{M} \quad (12)$$

and we are interested in estimating the depth values in \mathbf{M} , and \mathbf{A} is the matrix representing the observation process (rotation, translation and scaling). The estimation of \mathbf{M} from \mathbf{q} given \mathbf{A} can be viewed as a linear inverse problem. To solve the linear inverse problems one approach is to define a suitable cost function $\mathbf{J}(\mathbf{M})$ to find the global minimum. For this task we define the cost function as the sum of two terms and is given below:

$$\mathbf{J}(\mathbf{M}) = \mathbf{D}(\mathbf{q}, \mathbf{AM}) + \lambda \mathbf{R}(\mathbf{M}) \quad (13)$$

The first term $\mathbf{D}(\mathbf{q}, \mathbf{AM})$ measures the discrepancy between \mathbf{q} and \mathbf{AM} and $\mathbf{R}(\mathbf{M})$ is a regularization term. The parameter λ is called the regularization parameter and is used to adjust the trade-off between the two terms; λ should be a positive value. We will use the mean square error for $\mathbf{D}(\mathbf{q}, \mathbf{AM})$, namely

$$D(q, AM) = \|q - AM\|_2^2 \quad (14)$$

Minimizing this $D(q, AM)$ will give a signal M which is as consistent with q as possible according to the square error criterion. We could try to minimize $D(q, AM)$ by setting $M = A^{-1}q$; however, A may not be invertible. Even if A were invertible, it may be very ill-conditioned and the estimates may not represent the solution. On the other hand to penalize the undesirable behavior in M the regularization term $R(M)$ is used. As the signal M is a non-Gaussian, we choose l_1 -norm of matrix M as the regularization term, which is defined as

$$\|M\|_1 = \begin{matrix} \|M_{x1}\| \\ \|M_{y1}\| \\ \|M_{z1}\| \end{matrix} + \begin{matrix} \|M_{x2}\| \\ \|M_{y2}\| \\ \|M_{z2}\| \end{matrix} + \begin{matrix} \|M_{x3}\| \\ \|M_{y3}\| \\ \|M_{z3}\| \end{matrix} + \dots + \begin{matrix} \|M_{xn}\| \\ \|M_{yn}\| \\ \|M_{zn}\| \end{matrix} \quad (15)$$

Hence, the approach is to estimate M_z from q by minimizing the objective function

$$J(M) = \|q - AM\|_2^2 + \lambda \|M\|_1 \quad (16)$$

We have formulated the depth estimation problem as a l_1 -norm regularized linear inverse problem. In the literature the general approach to minimize l_1 -norm linear inverse problems Majorization-Minimization (MM) approach is used. This algorithm in general form appeared in optimization literature [25, 26] for processing 1-D signals and we propose a humble contribution of extending in the similar way to 2-D to extract the 3D face shape.

C. Majorization-Minimization

The MM can be described as follows. Suppose we have a vector M^k , a guess for the minimum of $J(M)$. Based on M^k , we would like to find a new M^{k+1} which further decreases $J(M)$; i.e., we want to find M^{k+1} such that $J(M^{k+1}) < J(M^k)$.

The MM approach asks us first to choose a new function which majorizes $J(M)$ and, second, that we minimize the new function to get M^{k+1} . MM puts some requirements on this new function, call it $G(M)$. We should choose $G(M)$ such that $G(M) \geq J(M)$ for all M . In addition $G(M)$ should equal $J(M)$ at M^k . We find M^{k+1} by minimizing $G(M)$. The function $G(M)$ will be different at each iteration. So we denote it $G_k(M)$. Applying the

MM procedure with l_1 -norm for estimating the 3D face structure results in iterative update equation and is given below:

$$M^{k+1} = \text{soft}(M^k + \frac{1}{\alpha} A^T (q - AM^k), \frac{\lambda}{2\alpha}) \quad (17)$$

$$\alpha \geq \max \text{eig}(A^T A)$$

The soft-threshold rule is the non-linear function defined

$$\text{as } \text{soft}(x, T) := \begin{cases} x + T & x \leq -T \\ 0 & |x| \leq T \\ x - T & x \geq T \end{cases} \quad (18)$$

III. EXPERIMENTAL RESULTS ON BOSPHORUS DATABASE

The first 30 subjects from the Bosphorus database [29] were used in the experiments. Note that images with unseen feature points cannot be selected as training images, as the corresponding depth values cannot be estimated. As a result, only five non-frontal-view face images, PR_D, PR_SD, PR_SU, PR_U and YR_R10 can be used to train the model in the experiments and are shown in Fig. 4 for one subject in the database. In the experiments, the pose parameters ϕ, ψ , and θ are initially set to be zeros. The scale parameter s is set to be 1. We can obtain one set of depth values for the facial-feature points when each non-frontal-view face image is combined with its corresponding frontal-view face image for the hybrid optimization. We have used DE strategy DE/rand/2 with Cr=0.2, F=0.4 and initial population of 30 and ISTA with $\lambda=0.0001$. All the simulations were conducted using MATLAB. The Pearson Linear correlation [30] coefficients of the true depth values and the estimated depth values of subjects 22-31 are given in Table I.

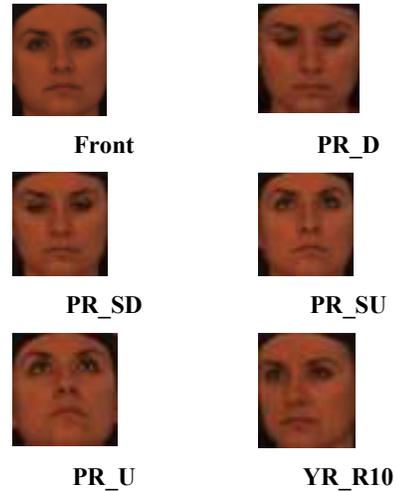


Fig.4 Face Images of one subject under different poses in the Bosphorus Database.

TABLE I. PEARSON CORRELATION COEFFICIENTS OF ESTIMATED DEPTH VALUES AND TRUE DEPTH VALUES OF SUBJECTS [22-31] OF BOSHPORUS DATABASE

	Bosphorus Database Training Images				
	PR_D	PR_SD	PR_SU	PR_U	YR_R10
Subj22	0.7550	0.7499	0.7567	0.7547	0.7532
Subj23	0.9119	0.8930	0.8766	0.8790	0.8822
Subj24	0.9293	0.9248	0.9209	0.9234	0.9226
Subj25	0.8907	0.8898	0.8777	0.8894	0.8838
Subj26	0.8645	0.8652	0.8476	0.8536	0.8626
Subj27	0.7813	0.7779	0.7743	0.7756	0.7685
Subj28	0.8798	0.8754	0.8727	0.8771	0.8844
Subj29	0.9214	0.9174	0.9280	0.9273	0.9328
Subj30	0.9104	0.9073	0.9083	0.9120	0.9145
Subj31	0.9453	0.9370	0.9345	0.9377	0.9308
GA_Subj1	0.1962	0.2787	0.5568	0.7283	0.5128

Further, Fig.5 shows the comparison of the correlation coefficients of the estimated depth values and Candide depths, estimated depths and true depth values and True depth values and candide depth values with one non-frontal-view face image and corresponding frontal view for the 31 subjects of the Bosphorus database.

Fig.6 shows the true depth values and estimated depths normalized to the interval [0,1] using proposed algorithm. Fig.6 it can be seen that depth values of most facial feature points are correctly estimated. It should be stressed that no prior information about the true depth values are used in the optimization procedure. The computation time required for GA and Proposed algorithm are given in Table. II.

TABLE II. COMPARISON OF COMPUTATIONAL TIME - ONE SUBJECT

GENETIC ALGORITHM	PROPOSED ALGORITHM
~50 SEC.	~10SEC.

IV. CONCLUSION

In this paper, efficient Optimization scheme is proposed to estimate the 3D face structure from 2D images. The proposed optimization scheme is comprised of Differential Evolution and followed by Iterative Soft Thresholding Algorithm. The experimental results verify that the proposed optimization scheme can improve the 3D face reconstruction accuracy. Compared to the GA based method, the proposed method has a comparable reconstruction performance, while the training time decreased significantly. Experimental results have demonstrated the accuracy of estimation consistently for majority of features. In future work, we will further improve the efficiency of the accuracy of the depth values estimated.

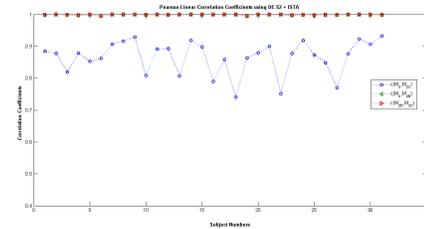


Fig.5. Pearson Correlation Coefficient obtained using PR_D as training sample .

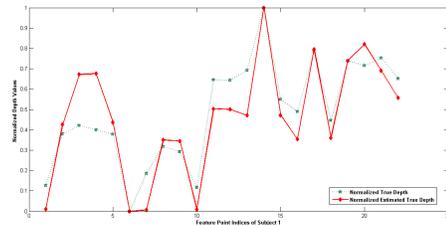


Fig.6. Comparison of True Depth and Estimated Depth Values of Subject 1.

REFERENCES

- [1] Chellappa, Rama, Charles L. Wilson, and Saad Sirohey. "Human and machine recognition of faces: A survey." *Proceedings of the IEEE* 83.5 (1995): 705-741.
- [2] He, Xiaofei, et al. "Face recognition using Laplacianfaces." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27.3 (2005): 328-340.
- [3] Bowyer, Kevin W., Kyong Chang, and Patrick Flynn. "A survey of approaches and challenges in 3D and multi-modal 3D+ 2D face recognition." *Computer vision and image understanding* 101.1 (2006): 1-15.
- [4] Levine, Martin D., and Yingfeng Chris Yu. "State-of-the-art of 3D facial reconstruction methods for face recognition based on a single 2D training image per person." *Pattern Recognition Letters* 30.10 (2009): 908-913.
- [5] Chowdhury, Amit Roy, and Rama Chellappa. "Statistical error propagation in 3d modeling from monocular video." *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*. Vol. 8. IEEE, 2003.
- [6] Koo, Hei-Sheung, and Kin-Man Lam. "Recovering the 3D shape and poses of face images based on the similarity transform." *Pattern Recognition Letters* 29.6 (2008): 712-723.
- [7] Thelen, Andrea, et al. "Improvements in shape-from-focus for holographic reconstructions with regard to focus operators, neighborhood-size, and height value interpolation." *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society* 18.1 (2009): 151-157.
- [8] Castelán, Mario, and Edwin R. Hancock. "Acquiring height data from a single image of a face using local shape indicators." *Computer Vision and Image Understanding* 103.1 (2006): 64-79.
- [9] Castelan, Mario, William AP Smith, and Edwin R. Hancock. "A coupled statistical model for face shape recovery from brightness images." *Image Processing, IEEE Transactions on* 16.4 (2007): 1139-1151.
- [10] Jiang, Dalong, et al. "Efficient 3D reconstruction for face recognition." *Pattern Recognition* 38.6 (2005): 787-798.
- [11] Romdhani, Sami, and Thomas Vetter. "Efficient, robust and accurate fitting of a 3D morphable model." *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003.

- [12] Zhang, Chongzhen, and Fernand S. Cohen. "3-D face structure extraction and recognition from images using 3-D morphing and distance mapping." *Image Processing, IEEE Transactions on* 11.11 (2002): 1249-1259.
- [13] Tomasi, Carlo, and Takeo Kanade. "Shape and motion from image streams under orthography: a factorization method." *International Journal of Computer Vision* 9.2 (1992): 137-154.
- [14] Fortuna, Jeff, and Aleix M. Martinez. "Rigid structure from motion from a blind source separation perspective." *International journal of computer vision* 88.3 (2010): 404-424.
- [15] Zhang, Ruo, et al. "Shape-from-shading: a survey." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 21.8 (1999): 690-706.
- [16] Ullman, Shimon. *The interpretation of visual motion*. Massachusetts Inst of Technology Pr, 1979.
- [17] Xirouhakis, Yiannis, and Anastasios Delopoulos. "Least squares estimation of 3D shape and motion of rigid objects from their orthographic projections." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.4 (2000): 393-399.
- [18] Bregler, Christoph, Aaron Hertzmann, and Henning Biermann. "Recovering non-rigid 3D shape from image streams." *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*. Vol. 2. IEEE, 2000.
- [19] Torresani, Lorenzo, Aaron Hertzmann, and Christoph Bregler. "Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30.5 (2008): 878-892.
- [20] Ahlberg, Jörgen. "An active model for facial feature tracking." *EURASIP Journal on applied signal processing* 2002.1 (2002): 566-571.
- [21] Xie, Xudong, and Kin-Man Lam. "Elastic shape-texture matching for human face recognition." *Pattern Recognition* 41.1 (2008): 396-405.
- [22] Storn, Rainer, and Kenneth Price. "Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces." *Journal of global optimization* 11.4 (1997): 341-359.
- [23] Das, Swagatam, and Ponnuthurai Nagaratnam Suganthan. "Differential evolution: a survey of the state-of-the-art." *Evolutionary Computation, IEEE Transactions on* 15.1 (2011): 4-31.
- [24] Beck, Amir, and Marc Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems." *SIAM Journal on Imaging Sciences* 2.1 (2009): 183-202.
- [25] Daubechies, Ingrid, Michel Defrise, and Christine De Mol. "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint." *Communications on pure and applied mathematics* 57.11 (2004): 1413-1457.
- [26] Figueiredo, Mário AT, and Robert D. Nowak. "An EM algorithm for wavelet-based image restoration." *Image Processing, IEEE Transactions on* 12.8 (2003): 906-916.
- [27] Figueiredo, Mário AT, José M. Bioucas-Dias, and Robert D. Nowak. "Majorization–minimization algorithms for wavelet-based image restoration." *Image Processing, IEEE Transactions on* 16.12 (2007): 2980-2991.
- [28] Selesnick, Ivan W. "Sparse signal restoration." Proceedings available online at url: <http://cnx.org/content/m32168/latest> (2010).
- [29] Savran, Arman, et al. "Bosphorus database for 3D face analysis." *Biometrics and Identity Management*. Springer Berlin Heidelberg, 2008. 47-56.
- [30] Hansen, Nikolaus, and Andreas Ostermeier. "Completely derandomized self-adaptation in evolution strategies." *Evolutionary computation* 9.2 (2001): 159-195.

Multi-lane traffic flow models accounting for different lane changing motivations

M. N. Smirnova, D. A. Pestov, A. I. Bogdanova, N. N. Smirnov, A. B. Kiselev, V. F. Nikitin and V. V. Tyurenkova

Abstract—The present research was aimed at mathematical modeling of essentially unsteady-state traffic flows on multilane roads, wherein The present research was aimed at mathematical modeling of essentially unsteady-state traffic flows on multilane roads, wherein massive changing of lanes produces an effect on handling capacity of the road segment. The model takes into account drivers' motivations for lane changing before the crossing caused by the necessity of the maneuver on entering multilane road crossing. The model is based on continua approach. However, it has no analogue with the classical hydrodynamics because momentum equations in the direction of a flow and in orthogonal directions of lane-changing are different. Thus velocity of small disturbances propagation in the traffic flow is different depending on direction: counter flow, co-flow, orthogonal to the flow. Numerical simulations of traffic flows in multilane roads were performed and their results are presented

Keywords—Traffic, flow, control, continua model, multilane road, handling capacity.

I. INTRODUCTION

FIRST attempts of developing the mathematical models for traffic flows were undertaken by Lighthill and Whitham [1], Richards [2], Greenberg [3], Prigogine [4]. A detailed analysis of the results obtained could be found in the book by Whitham [5]. The authors regarded the traffic flow from the position of continuum mechanics applying empirical relationships between the flow density and velocity of vehicles. Development of computer technique gave birth to continua models application not only for analytical solutions

M. N. Smirnova is with the Moscow M.V. Lomonosov State University, Moscow 119992, Russia and Saint Petersburg State Polytechnical University, 29 Politechnicheskaya Str., St. Petersburg 195251, Russia (email: wonrims@inbox.ru)

D. A. Pestov is with the Moscow M.V. Lomonosov State University, Moscow 119992, Russia

A. I. Bogdanova is with the Moscow M.V. Lomonosov State University, Moscow 119992, Russia (email: sashabogdanova@gmail.com)

N. N. Smirnov is with the Moscow M.V. Lomonosov State University, Moscow 119992, Russia (email: ebifsun1@mech.math.msu.su)

A. B. Kiselev is with the Moscow M.V. Lomonosov State University, Moscow 119992, Russia (email: akis2006@yandex.ru)

V. F. Nikitin is with the Moscow M.V. Lomonosov State University, Moscow 119992, Russia (email: vfnikster@gmail.com)

V. V. Tyurenkova is with the Moscow M.V. Lomonosov State University, Moscow 119992, Russia (corresponding author to provide phone: +79166236736, email: turenkova.v.v@rambler.ru)

[6] but also in numerical simulations of traffic flows [7-10]. Another approach using discrete approximation for public traffic and passengers interaction is illustrated by [11-12].

The present research was motivated by the arising difficulties in rational traffic organization the Moscow Government was faced in the recent years. The decrease of the roads handling capacity in the centre of the city caused by the increase of traffic density brought to necessity of re-organizing the traffic flows and constructing new roads in and around the city.

II. THE CONTINUA MODEL OF TRAFFIC FLOWS.

We introduce the Euler's co-ordinate system with the Ox axis directed along the auto route and the Oy axis directed across the traffic flow. Time is denoted by t . The average flow density $\rho(x, y, t)$ is defined as the relation of the surface of the road occupied by vehicles to the total surface of the road considered.

$$\rho = \frac{S_{tr}}{S} = \frac{hn\ell}{hL} = \frac{n\ell}{L},$$

where h is the lane width, L is the sample road length, ℓ is an average vehicle's length plus a minimal distance between jammed vehicles, n is the number of vehicles on the road. With this definition, the density is dimensionless changing from zero to unit.

The flow velocity denoted $V(x, y, t) = (u(x, y, t), v(x, y, t))$, where u can vary from zero to U_{max} , where U_{max} is maximal permitted road velocity. From definitions, it follows that the maximal density $\rho = 1$ relates to the case when vehicles stay bumper to bumper. It is naturally to assume that the traffic jam with $V = 0$ will take place in this case.

Determining the "mass" distributed on a road sample of the area S as:

$$m = \int_S \rho d\sigma,$$

one can develop a "mass conservation law" in the form of continuity equation:

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho u)}{\partial x} + \frac{\partial(\rho v)}{\partial y} = 0. \quad (1)$$

Then, we derive the equation for the traffic dynamics. The traffic flow is determined by different factors: drivers reaction on the road situation, drivers activity and vehicles response, technical features of the vehicles. The following assumptions were made in order to develop the model:

- 1) It is the average motion of the traffic described and not the motion of the individual vehicles, that is modeled. Consequently, the model deals with the mean features of the vehicles not accounting for variety in power, inertia, deceleration way length, etc.
- 2) The “natural” reaction of all the drivers is assumed. For example, if a driver sees red lights, or a velocity limitation sign, or a traffic hump ahead, he is expected to decelerate until full stop or until reaching a safe velocity, and not to keep accelerating with further emergency braking.
- 3) It is assumed that the drivers are loyal to the traffic rules. In particular, they accept the velocity limitation regime and try to maintain the safe distance depending on velocity.

The velocity equation for the x-component is then written as follows:

$$\frac{du}{dt} = a; \quad a = \max \left\{ -a^-; \min \{ a^+; a' \} \right\};$$

$$a' = \sigma a_\rho + (1 - \sigma) \int_0^y \omega(y) a_\rho(t, s) ds + \frac{U(\rho) - u}{\tau}$$

$$a_\rho = -\frac{k^2}{\rho} \frac{\partial \rho}{\partial x},$$

Here, a is the acceleration of the traffic flow; a^+ is the maximal positive acceleration, a^- is the emergency braking deceleration; a^+ and a^- are positive parameters which are determined by technical features of the vehicle. The parameter $k > 0$ is the small disturbances propagation velocity (“sound velocity”), as it was shown in [16-18]. The parameter τ is the delay time which depends on the finite time of a driver’s reaction on the road situation and the vehicle’s response. This parameter is responsible for the drivers tendency to keep the vehicles velocity as close as possible to the safe velocity depending on the traffic density $U(\rho)$ [16-18]:

$$U(\rho) = \begin{cases} -k \ln \rho, & u < U_{\max}, \\ U_{\max}, & u \geq U_{\max}. \end{cases}$$

The velocity $U(\rho)$ is determined from the dependence of the traffic velocity on density in the “plane wave” when the traffic is starting from the initial conditions $\rho_0 = 1$ and $u = 0$, with account of velocity limitation from above ($u \leq U_{\max}$). The value of τ could be different for the

cases of acceleration or deceleration to the safe velocity $U(\rho)$:

$$\tau = \begin{cases} \tau^+, & U(\rho) < u \\ \tau^-, & U(\rho) \geq u \end{cases}$$

In the formula for a' the first term describes an influence on the driver’s reaction in a local situation, the second – a situation ahead the flow and the third – a driver’s tendency to drive a car with the velocity which is the safest in each case. If we assume $\sigma = 0$, then the expression for a' will be:

$$a' = \frac{1}{\Delta} \int_x^{x+\Delta} a_\rho(t, s) ds + \frac{U(\rho) - u}{\tau}$$

In this case acceleration is not a local parameter, but depends on its values in the region of length Δ ahead of vehicle, where Δ - is the distance each driver takes into account on making decisions. This distance depends on road and weather conditions. The last term of the relaxation type takes into account tendency to reach optimal velocity.

If we assume $\sigma = 1$, then the expression for a' will be:

$$a' = a_\rho + \frac{U(\rho) - u}{\tau}$$

Now acceleration depends on local situation only.

We will consider a case when $\sigma = 1, \tau = \infty$, so the equation of motion for the x-component is:

$$a_x = -\frac{k^2}{\rho} \frac{\partial \rho}{\partial x}. \quad (2)$$

The equation of motion for the y-component we can write in such form as for the x direction:

$$a_y = -\frac{A^2}{\rho} \frac{\partial \rho}{\partial x}. \quad (3)$$

The description for the parameter A will be given below.

In order to understand the physical meaning of parameter A we shall consider the following model problem. One car is changing it’s lane with the density $\rho \neq 0$ to the lane with the density $\rho = 0$. In this case a car has the maximum

acceleration a_{\max} . So by putting $\frac{\partial \rho}{\partial y} = \frac{0 - \rho}{h}$ to the velocity

equation for the y-component we can obtain

$$\rho \frac{dv}{dt} = -A^2 \frac{\partial \rho}{\partial y} \text{ and then } \frac{A^2}{h} = a_{\max}.$$

The diagram for the $v(y)$ is showed on the Fig.1.

According to it we have: $v_{\max} = \sqrt{a_{\max} h}$, that is

$$a_{\max} = \frac{v_{\max}^2}{h} \text{ or } A^2 = v_{\max}^2. \text{ But in this model}$$

$v_{\max} = 2v_{av}$ (v_{av} is an average speed of car's changing a lane), so $A^2 = 4v_{av}^2$, where v_{av}^{\max} - the average speed for V_{\max} .

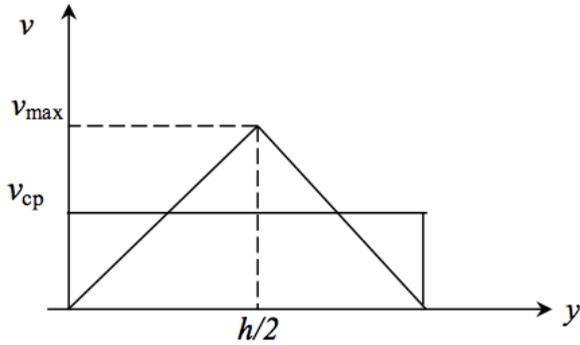


Fig. 1 the diagram for the $v(y)$.

For description of lane change dynamics we approximately assume that the trajectory of maneuver of lane's changing is assembled of two parts of identical circles (Fig. 2). On the Fig.2 the bold line is a trajectory of lane's changing and it begins in the middle of the lane on which the car is going and ends in the middle of next lane.

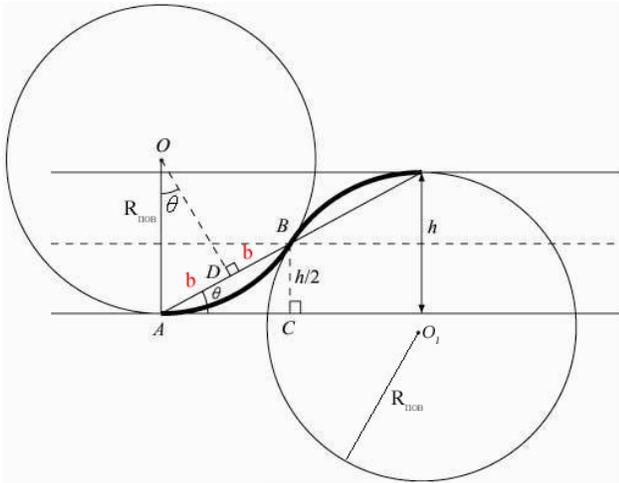


Fig. 2 the trajectory of maneuver of lane's changing.

The centripetal force which acts on the car is: $F = \frac{mV^2}{R_{turn}}$,

where R_{turn} is the radius of the turn, V - is the velocity of the car, directed by the tangent to the car's trajectory. Let F_* be the maximum possible flank force under which car is drivable (not skidding), $F \leq F_*$. Then we may calculate the radius of the

turn $R_{turn} = \frac{m}{F_*} V^2 \cdot v_{av}^{\max} = V \sin \theta$. From the similar

triangles $\triangle ABC$ and $\triangle OAD$ (Fig. 2) derive that $AD = DB$, labeling $AD = b$ will get:

$$\frac{b}{R} = \frac{h/4}{b} \text{ or } b^2 = \frac{Rh}{4},$$

$$\sin \theta = \frac{h}{4b} = \frac{b}{R} = \frac{\sqrt{Rh}}{2R} = \frac{1}{2} \sqrt{\frac{h}{R}}.$$

$$\text{Then } v_{av}^{\max} = V \sin \theta = \frac{1}{2} V \sqrt{\frac{h}{R}} = \frac{1}{2} V \sqrt{\frac{hF_*}{mV^2}} = \frac{1}{2} \sqrt{\frac{hF_*}{m}}$$

that is the average speed for V_{\max} is $v_{av}^{\max} = \frac{1}{2} \sqrt{\frac{hF_*}{m}}$. We

obtained above $A^2 = 4v_{av}^2$, so the expression for A is

$$A^2 = \frac{hF_*}{m}.$$

Thus we derived parameter A^2 being dependent on the force F_* , assignable to the lane width, mass of the car and the traction of tires.

The equations (1), (2), (3) provide a system describing traffic flows on multilane roads. Assuming the presence of 3 groups of cars before each crossing characterized by different motivations: going straight, turning left and turning right, - one can describe their motivations by introducing different mass forces acting on each group of cars and making them move towards right, central or left lanes, as well as slow down the speed of their motion on approaching the crossing.

$$\frac{\partial \rho_i}{\partial t} + \frac{\partial (\rho u)_i}{\partial x} + \frac{\partial (\rho v)_i}{\partial y} = 0,$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = -\frac{k^2}{\rho} \frac{\partial \rho}{\partial x} + g,$$

$$\frac{\partial v_i}{\partial t} + u \frac{\partial v_i}{\partial x} + v_i \frac{\partial v_i}{\partial y} = -\frac{A^2}{\rho} \frac{\partial \rho}{\partial y} + f_i;$$

$$i = 1, 2, 3; \rho = \rho_1 + \rho_2 + \rho_3$$

$$v = \frac{1}{\rho} (\rho_1 v_1 + \rho_2 v_2 + \rho_3 v_3)$$

The force slowing down speed before entering crossing has the following model form:

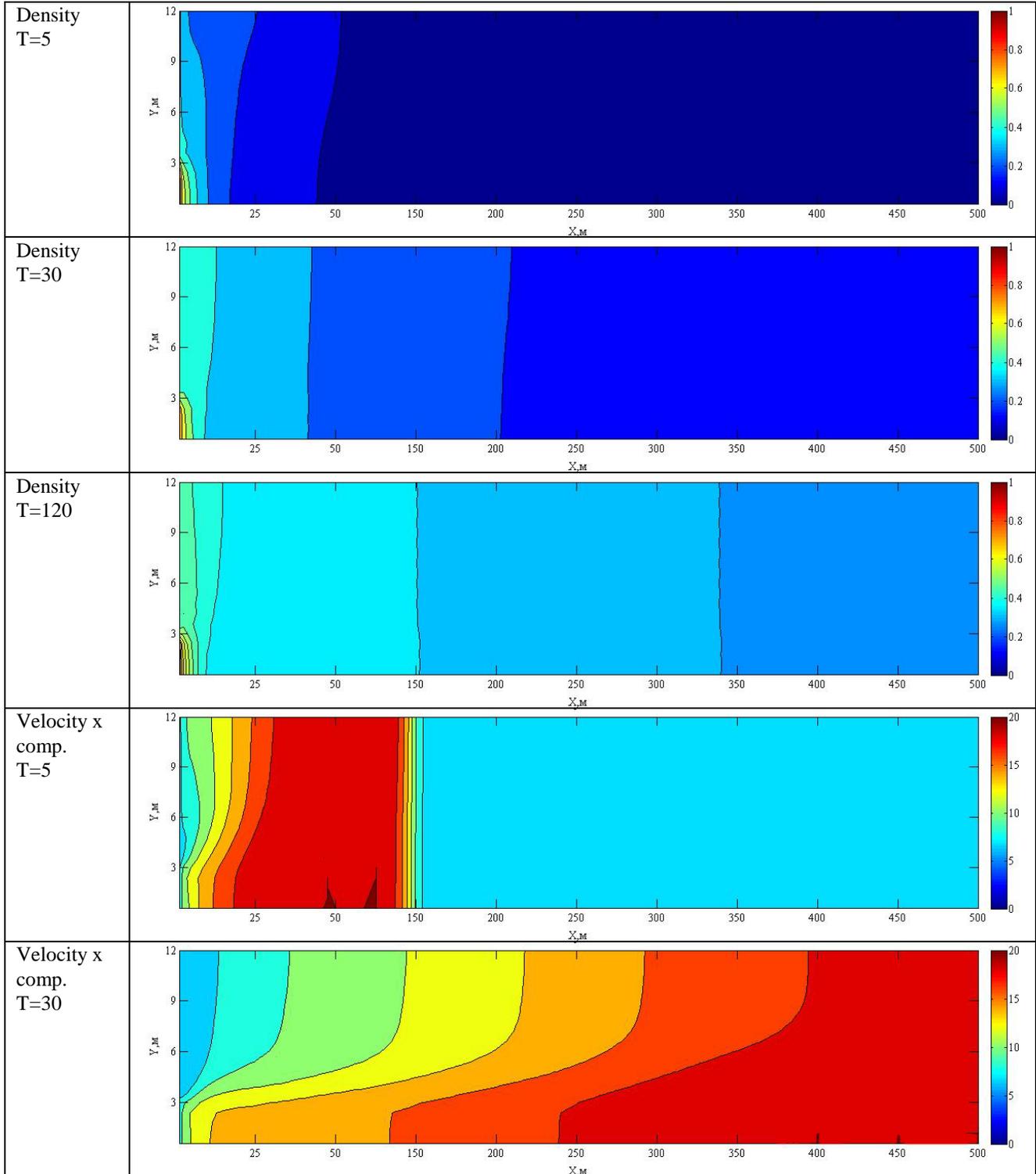
$$g = -(1-\rho) \frac{U(t)-u}{\tau_0} \exp\left(-\frac{x_0-x}{l_s}\right), \text{ where } l_s \text{ is the}$$

characteristic distance of beginning deceleration in front of the crossing and traffic lights, τ_0 - characteristic time of reaction on traffic light change.

The forces responsible for the modeling motivation for lane changing could be expressed as follows:

$$f_i = \begin{cases} -(1-\rho_i) \frac{A^2}{h_i} \exp\left[-\frac{x_0-x}{l_m}\right], & y > y_i^+ \\ 0, & y_i^+ > y > y_i^- \\ (1-\rho_i) \frac{A^2}{h_i} \exp\left[-\frac{x_0-x}{l_m}\right], & y < y_i^- \end{cases}$$

$y_1^- = 0, y_1^+ = y_2^- = \frac{H}{3}, y_2^+ = y_3^- = \frac{2H}{3}, y_3^+ = H$
 where h_i is the respective lane width, H – the road width.



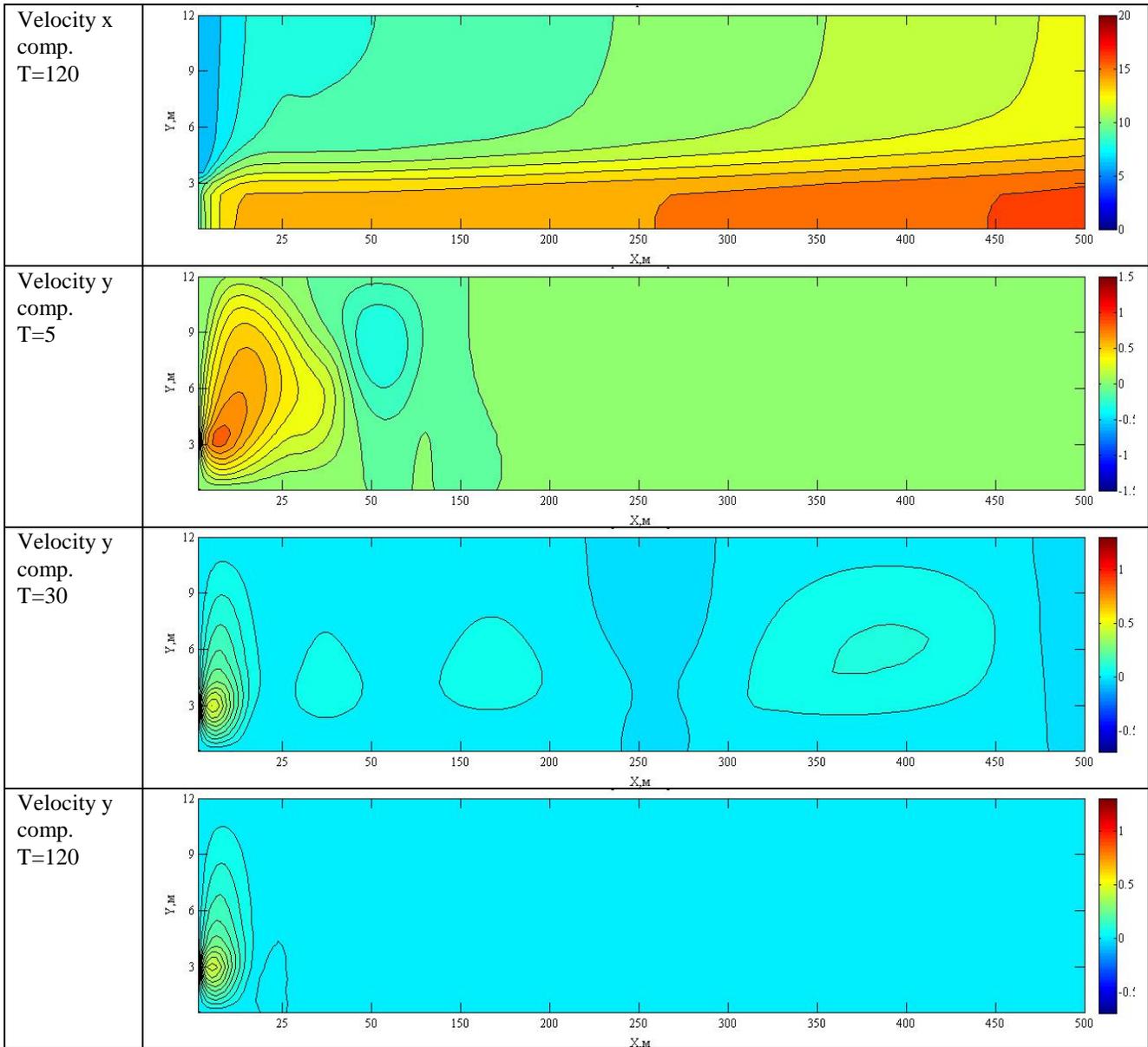


Fig. 3 the maps for components of velocity and for density for different moments of time.

III. SOLVING A TEST PROBLEM.

The considered problem is the nature of car's behavior on a multilane rectangular road.

The traffic flow is divided into two parts on the left boundary by y_{fix} . For $0 \leq y \leq y_{fix}$ the flow has the bigger value then for $y_{fix} < y \leq H$.

Boundary conditions for $x=0$:

$$\left. \begin{aligned} v = 0, u = u_1, \rho = \rho_1, u_1 \rho_1 = q_1 \\ v = 0, u = u_2, \rho = \rho_2, u_2 \rho_2 = q_2 \end{aligned} \right\} u > k \quad \text{and}$$

$$q = \rho u = \begin{cases} q_1, & y > y_{fix} > 0 \\ q_2, & y \leq y_{fix} \leq H \end{cases}$$

Boundary conditions for $x=L$:

$$\frac{\partial \rho u}{\partial x} = 0, \frac{\partial \rho v}{\partial x} = 0,$$

The numerical calculations of the problems were processed using the AUSM method. The mesh had $N_x = 201$ and $N_y = 21$ grid nodes.

$T = 240$ s - calculation time;

$L = 500$ m - the length of the domain;

$H = 12$ m – the width of the domain;

$\rho_0 = 0,01$ – the density of traffic flow on the boundary $x = 0$;

$u_0 = 7,0$ m/s – the x-velocity of traffic flow on the boundary $x = 0$;

$v_0 = 0,0$ m/s – the y-velocity of traffic flow on the boundary $x = 0$;

$U_{\max} = 20$ m/s – maximal permitted road velocity;

$k = 9$ m/s – the small disturbances propagation velocity (“sound velocity”) on x axis;

$A = 3$ m/s – the analogy of k coefficient on y axis;

$y_{fix} = 3,75$ m – the divisor of the road on 2 parts;

$q_0 = 7$ – a flow on the left boundary for $0 \leq y \leq y_{fix}$;

$q_0 = 3$ – a flow on the left boundary for $y_{fix} < y \leq H$;

The obtained results are presented in the form of maps for components of velocity and for density for different moments of time (Fig.3).

IV. CONCLUSIONS

The mathematical model for traffic flows simulations in multi-lane roads is developed. A model problem for traffic evolution in multi-lane road with non-uniform flux in different lanes was regarded. The results show that on entering the road segment high orthogonal fluxes occur in the direction of less dense lanes, which brings to slowing down flow velocity in that lanes and increasing density. The equation of motion at brings to formation of inverse flow from those lanes back. In time flow evolution brings to a more smooth transition zone, however, on coming in contact flow zones of different fluxes always cause formation of an orthogonal flow in the direction of less density lane in the nearest vicinity followed by an inverse flow at some distance.

ACKNOWLEDGEMENT

Russian Foundation for Basic Research (grant 13-01-12056) is acknowledged for financial support.

REFERENCES

- [1] M. J. Lighthill, J. Whitham, On kinetic waves II. A theory of traffic flow on long crowded roads. *Proc. of the Royal Society. Ser. A. No. 1178. Vol. 229. London, 1955. 317-345.*
- [2] P. L. Richards, Shock waves on the highway. *Operations Research. 1956. Vol. 4. 42-51.*
- [3] H. Greenberg, An analysis of traffic flow. *Operations Research. 1959. Vol. 7. 79-85.*
- [4] I. Prigogine, P. Résibois, On a generalized Boltzman-like approach for traffic flow. *Bull. Cl. Sci., Acad. Roy. Belg., 1962, vol. 48, No. 9, 805-814.*
- [5] J. Whitham, *Linear and non-linear waves.* Moscow, John Wiley & Sons, New York – London, 1974.
- [6] G. P. Soldatov, On formation of a strong shock wave in a two-directional traffic flow. *Applied Math. & Mech., 1970, vol. 34, No. 1, 135-137.*
- [7] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, Y. Sugiyama. Dynamical Model of Traffic Congestion and Numerical Simulation, *Physical Review. 1995. Ser. E. Vol. 51. P. 1035.*
- [8] S. A. Hill Numerical analysis of a time-headway bus route model, *Physica A. 2003. Vol. 328, No. 1-2. P.261-273.*
- [9] Md. M. Karim, T. Ohno, Air quality planning and empirical model to evaluate SPM concentrations, *Journal of Environmental Engineering. 2000. Vol. 64. P. 1116-1124.*
- [10] A. B. Sukhinova, M. A. Trapeznikova, B. N. Chetverushkin, N. G. Churbanova, Two dimensional macroscopic model for traffic flows. *Mathematical modeling. 2009 vol. 21, №2, pp.118-126.*
- [11] T. Nagatani Bunching transition in a time-headway model of a bus route, *Phys. Rev. E. 2001. Vol. 296, № 1-2. P.320-330.*
- [12] S. A. Regirer, N. N. Smirnov, A. E. Chenchik, Mathematical model of moving collectives interaction: Public transport and passengers. *Automation and Remote Control. 2007, vol. 68, No 7, pp. 1225-1238.*
- [13] A. B. Kiselev, A. V. Kokoreva, V. F. Nikitin, N. N. Smirnov, Computational modelling of traffic flows, *European Conf. on Computational Fluid Dynamics* (Egmond aan Zee, The Netherlands, 5-8 Sept. 2006). Book of Abstracts. – Delft: TU Delft, 2006, p. 265.
- [14] A. B. Kiselev, A. V. Kokoreva, V. F. Nikitin, N. N. Smirnov, Computational modelling of traffic flows, *European Conf. on Computational Fluid Dynamics* (Egmond aan Zee, The Netherlands, 5-8 Sept. 2006). Proc. on CD-Rom. – 10 p.
- [15] N. N. Smirnov, A. B. Kiselev, V. F. Nikitin, M. V. Yumashev, Mathematical modeling of traffic flows. *Moscow University Mechanics Bulletin, 2000, № 4, pp. 39–44.*
- [16] A. B. Kiselev, A. V. Kokoreva, V. F. Nikitin, N. N. Smirnov, Mathematical modeling of traffic flow on traffic-controlled roads. *Journal of Applied Mathematics and Mechanics* (Engl. Transl.) 2004, Vol.68, No 5, pp. 1035–1042.
- [17] A. B. Kiselev, A. V. Kokoreva, V. F. Nikitin, N. N. Smirnov, Mathematical modeling of traffic flow dynamics. *Proc. of M. V. Lomonosov Conf. 2003, Moscow, p. 70.*
- [18] N.N. Smirnov, A.B. Kiselev, V.F. Nikitin, M.V. Yumashev, Mathematical modelling of traffic flows. *Proc. 9th IFAC Symposium Control in Transportation Systems 2000, Braunschweig, 2000.*
- [19] A. B. Kiselev, V. F. Nikitin, N. N. Smirnov, M. V. Yumashev, Irregular traffic flow on a ring road. *Journal of Applied Mathematics and Mechanics* (Engl. Transl.) 2000, Vol.64, No 4, pp. 627-634.
- [20] N. N. Smirnov, A. B. Kiselev, V. F. Nikitin, A. V. Kokoreva Mathematical modeling of traffic flows using continua approach. Two lanes roads: modeling of T-shape crossing and the effect of lane changing on handling capacity. *Proceedings of MIPT, 2010, vol.2, N.4, pp.141-151.*

Potential of pervasive computing through embedded systems and Internet technologies: Research of customer perspective

Nikola Vlahović, Jovana Zoroja, Vesna Bosilj Vukšić

Abstract—During the recent economic recession software industry has shown a high level of resilience to negative economic trends by initiating a substantial innovation cycle. In turn, European commission recognized software industry and consumer electronics to be key industries that were able to change the course of economic trends. Most of these innovative endeavors relied on technologies that are the basis of pervasive computing, such as mobile applications, consumer electronics, embedded systems and Internet technologies. In order to create significant economic benefits it is not sufficient to have technological capacity to produce these innovations, but there are equally challenging issues of targeting the customer markets so that these innovations get accepted and absorbed by the market. Better understanding of how the demand for these services is created and what are its main features.

In this paper we will present the research of market potential for advanced services that combine best practices in embedded systems with Internet computing, its success in implementation through the perspective of users that are the targeted customers of these services.

Goal of the paper is to investigate users' understanding and awareness of the underlying technologies on one hand and to understand the level of adoption of currently available services as well as to estimate the potential of implementing future services based on pervasive computing concepts.

Results of the research show that there is a significant lack of understanding of these technologies by the end users. Nonetheless, this fact does not present an obstacle nor does it reduce the interest for using the services associated with these technologies.

Keywords—Consumer electronics markets, Embedded systems, Innovation, Internet technologies, Mobile computing, Pervasive computing, Ubiquitous computing.

I. INTRODUCTION

RECENT global economic crises has initiated a very successful technological innovation cycle that has introduces a variety of new products and services in the customer electronics, software services, mobile services and internet market segments. The industries that supply these market segments have presented a high level of resilience to negative economic trends. Even though the purchasing power has been decreased newly established customer habits as well as new economic and business models managed to alleviate negative trends and open new opportunities. One of most interesting segments that have been under development is the pervasive computing and its applications. These applications offer high level of convenience for their users, and make the devices that they already invested in even more useful. On the other hand there is a number of potential drawbacks and issues arising from the intensive use of the underlying technologies,

such as security and privacy issues.

In this paper we will take a look at recent technological developments from the customer perspective and try to understand how customers decide to accept new products and services offered through embedded and mobile systems that make up the backbone of future pervasive computing applications. Since pervasive computing creates applications and services that have additional level of autonomous operation many of technological facts about these technologies are hidden from users. It is fair to ask whether the lack of customer involvement in understanding the technology also creates additional incentive to use such products and services. On the other hand the high dynamics of contemporary life styles may be overwhelming for customers so that they may accept new trends, fads and services without assessing potential risks or even that the customers' level of tolerance to risk has grown due to thrust in new reliable technologies. It is also fair to ask where is the limit of this new thrust and what is the potential of new pervasive computing technologies in the future.

Goal of this paper is twofold: firstly, we want to estimate the level of understanding of recent technological developments by the individuals who are target audience for the products and services produced using these innovations. Secondly, we wish to estimate the potential interest in currently available services that fall under the pervasive computing paradigm as well as to investigate the interest of potential customers to new and innovative services that are currently either being developed or being designed. When we acquire this information we will be able to discuss the relation between the knowledge about technology and readiness of customers to use resulting products and services, which may serve as invaluable indicator for the development of future innovative projects in the underlying industries.

The structure of the paper is organized in six Sections. After the introduction in Section I, an overview of pervasive and ubiquitous computing is given in Section II. The emphasis is on the recent developments concerning primarily embedded systems, Internet technologies and mobile technologies. In Section III overview of the economic aspects and customer markets where products and services based on technologies and areas of pervasive computing will be presented. In the following Section, a research about the potential of embedded systems, mobile applications and Internet applications from the customer perspective will be presented. In Section V a

discussion of the research results will be given. Here the relation between the understanding of technologies and usage will be investigated. Finally in Section VI conclusions will be given as well as the remarks for future work.

II. OVERVIEW OF RECENT DEVELOPMENTS IN PERVASIVE COMPUTING

Pervasive or ubiquitous computing is a general term describing the new conceptual paradigm where most of the devices and their environment have particular computing capabilities. This term is meant to contrast the term desktop computing in the sense that computing is available from any device, any location using any computing aspect or technology.

Goal of pervasive computing is to provide users with proactive and self-tuning environments and devices that can augment personal knowledge and decision making abilities, while requiring as little direct user interaction as possible [1].

Pervasive paradigm relies primarily on embedded systems, mobile technologies and Internet technologies to supply everyday devices with computing capabilities. Devices include electronics that have microprocessors by definition, but also everyday objects or mechanical devices that traditionally did not contain computing capabilities.

Pervasive environment is based on creating hybrid ubiquitous networks resulting in augmented reality models based primarily on wireless technologies, mobile technologies in particular and internet technologies. These networks are highly dynamic such as Near Area Networks (NANs) or Body Area Networks (BANs), but also incorporate a reasonable degree of intelligence, allowing for autonomous decision making, self-organization [25] etc. Environment that can react to the presence of people is also referred to as the ambient intelligence.

Other technologies are also used, primarily RFID technology GPS technology and other sensor and actuator technologies.

Development of pervasive systems is an ongoing process resulting in a number of applications spanning from very simple control/indicator systems to more elaborate applications including mobile devices and electronic business models etc.

We will present some of the current research for future applications below. One of the most studied applications is the development of smart homes that use ambient intelligence to provide comfortable, assistive and secure environment for their residents [10, 13, 14]. Sentient computing systems that brings various services to the user using virtual networks such as redirecting phone calls to the phone nearest to the user or teleporting user's desktop to the nearest available computer, etc... [12]. Exploritorium project applied pervasive computing to museums and other exhibition venues augmenting the visitor experience by adding the additional layer of interaction with the exhibits [8]. For computation intensive scientific research a pervasive computing project e-Science was employed to automatically capture and publish data from ongoing experiments in laboratories [21]. One of few implementations that were active in the real world situations is the iHospital

system that supported the coordination of operations in a large hospital [11]. Other notable research examples include pervasive computing in automobiles [9, 20] and outdoor applications for both civic and military implementations [17].

In order to efficiently construct and use pervasive environments, devices should be enabled to perform basic computing tasks and participate in networking. In the rest of this section we will make an overview of the most important technologies used for this purpose, embedded systems and Internet and mobile technologies.

A. Embedded systems

Embedded systems are computer systems created to perform few very specific tasks. Due to their conceptual dedication to solving a particular set of problems and tasks they are not easy to change. They may include both software and hardware components, but they do not resemble general purpose computer systems. They are mostly added to mechanical or electronic devices as permanent part providing these devices with additional capabilities and improving their initial properties and efficiency. They are used to enable these devices to process information in real-time or process and exchange data with other systems. Embedded systems are used to control most of devices common to contemporaneity. Key part of all embedded systems is software component that cannot be changed once deployed within the device. Some devices though, have gone through a process of evolving from basic embedded systems towards more general purpose devices such as mobile phones and other mobile devices that may periodically update their software components called firmware.

Embedded systems have become commonly used in most application areas. The fact that 98% of all the microprocessors produced in the world use embedded systems [7] supports the importance of the development and evolution of embedded systems.

Due to their presence and improved capabilities embedded systems provide new solutions in the way that user may not perceive and be aware of them so the gap between the users active understanding and the rate of usage of these systems is increasing. This characteristic is also translated from embedded systems to pervasive computing where providing innovative and convenient new functionality does not require the user to understand the inner working [1]. While this may be an advantage in terms of user interaction and convenience, it may also pose a reasonable threat in terms of security, privacy and device failure scenarios.

B. Internet and mobile technologies

Internet usage is increasing every day because of its huge potential and benefits to the end users [6]. Internet technologies are changing our everyday life and they become ubiquitous in private and business life [19]. Development and adoption of Internet technology has positive impact on competitiveness and social-economic growth of countries, firms and individuals [4]. In 2014 there was 454.2% more Internet users in Europe than in 2000 year and worldwide

growth of Internet users from 2000 to 2014 year was 741.0% [16].

Mobile technologies are closely related to the Internet and they change the way individuals manipulate information and how they are connected with each other. Mobile technologies consist of hardware, software, and network infrastructure and offer wide spectrum of functionality to the users [5]. Mobile phones provide information regardless time and place constraints [22] and they offer the potential to different private and business activities. There are many mobile applications that come pre-installed or can be added to mobile phones, e.g. video, photography, high-speed internet access, social networking, Bluetooth, email, games, GPS location services [5, 22]. In many countries, more and more individuals use smart mobile phones more than laptop or personal computers. In the USA 58% of adults use mobile phones in order to check their electronic mail, communicate or manipulate with data.

There are many individuals who uses different Internet and mobile technologies and devices which include embedded system but who are not understanding their functions and do not use all possibilities they offer. In this article we want to estimate the level of understanding of technologies by individuals who use them at daily basis.

III. CUSTOMER MARKETS AND PERVASIVE COMPUTING CAPABILITIES

Innovation in ICT industry is feasible only if the end user market can absorb newly developed products and services. Same is true for goods and services rooted in pervasive computing. Understanding the current state of customer markets that are related to pervasive computing is of utmost importance. Here, we will take a closer look at the consumer electronic markets and their current states in terms of revenue and structure. Also we will take a closer look at the performance of mobile application distribution platforms that present a key component in distribution of current customized applications of various products and services that originate in embedded systems, mobile technologies or other networking technologies that are the basis of pervasive computing.

Consumer electronic market includes all the markets of electronic equipment used by individuals on daily basis either for personal or official use. Consumer electronic market contains several market segments: (1) digital media boxes, which include home video game consoles, blu-ray video players and recorders and digital media adapters or multimedia gateways; (2) computers, that include desktop computers and laptop computers, but also other sub segments such as netbooks and notebooks; (3) Televisions; (4) Set-top boxes, subdivided according to external signal technology such as cable, satellite, terrestrial or IPTV signal; (5) portable media devices that include sub segments such as smartphones, handheld game consoles and media tablets.

According to World Consumer Electronics Market 2014-2018 Research Report [15] the market is expected to grow further due to favorable conditions. The report further details

four most important influences which are active technological innovation cycle, availability of new services and positive response by the end users in adoption of new customer habits. Mobile devices are bestsellers worldwide. There is an increase in the sales of connectable devices, and there is a shift from stationary desktop computer sales towards more mobile technologies and devices such as tablet computers.

The sales of these products are mainly determined by the technology used in these products, the pricing, availability of different variants, and the level of after-sales support given to customers. In addition, global economic conditions also influence the sales of consumer electronics because they directly affect the purchasing power of customers.

During the recent years of global economic recession most of the segments of consumer markets showed high resilience to economic fluctuations, counteracting negative trends through technological innovation cycle that has been active since 2009.

This means that availability of innovative information devices has not decreased and that through everyday use individuals have already accepted new habit of everyday use which is also a positive indication for further innovations such as pervasive computing.

Closely related to customer electronics markets is the software market. Global software sales have not experienced a significant drop in revenues during the recent global economic crises [23]. Software industry trends are showing an increasing share of mobile application products being distributed to customer markets. The importance of this segment is also recognized by software developers and new software development methodologies are being employed in the development processes, e.g. Mobile-D [2, 18].

Mobile application distribution platforms play a major role in providing availability and accessibility of new mobile applications. Additionally they are a readily available distribution channel for after-sales support for customers that strongly influence the creation of new customer habits and expectations.

The revenue shares of products and services in mobile computing are steadily increasing since its first appearance. Recent study estimates that in the EU region revenues from mobile applications have reached over 10 billion euros in 2013, while creating over half a million jobs in 28 EU countries [24]. This segment has gained in economic significance as the market shares have risen, not only by the revenue attained through the realization of products and services but also by creating jobs, and establishing new service industries with related supporting service industries. This is why this market segment, taken as a whole, both the customer side and supplier side, is being referred to as the App Economy. The role of App Economy will be crucial in further development of products and services of the pervasive computing paradigm.

Takin into account the positive market trends in accepting new product and services and the fact that pervasive computing is eliminating the need to understand the underlying

technologies in these products and services we will present a research that tries to explain if there is the lack of understanding of technologies used today and how this may affect future trends from customer perspective.

IV. RESEARCH OF PERVASIVE COMPUTING POTENTIAL SUPPORTED BY EMBEDDED SYSTEMS AND INTERNET TECHNOLOGIES FROM THE CUSTOMER PERSPECTIVE

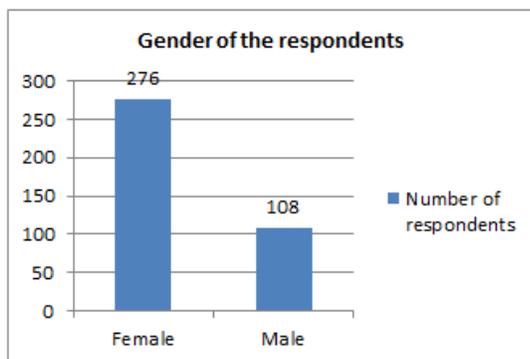
Pervasive computing through embedded systems and internet technologies is developing every day and its usage is growing constantly. However, there is still lack of research and studies dealing with this topic especially regarding users. In this article we want to estimate the level of understanding of using internet technologies by the individuals and to investigate the interest of individuals to new and innovative Internet technology services. In order to achieve these two goals a simple close-ended questionnaire was conducted.

A. Methodology and data

In May 2013 the empirical research on usage of ubiquitous computing supported by embedded systems and internet technologies was conducted among the third year students from the Faculty of Economics and Business Zagreb. Students were engaged in the course Business Information Systems. There were 384 students who participated in the survey.

The survey consists of two main parts. The first part of the survey is about embedded systems and how students are informed regarding their usage. In the second part of the survey participants named which of various devices and service which comprise embedded systems they use. While fulfilling the survey, the respondents can choose one or more offered answers. The questionnaire was distributed among students through mail in google docs format. The questionnaire was analyzed using descriptive statistics methods and techniques with the Statistica software.

Demographic characteristics of the respondents show that most of the students are female who are 21 year old. They are attending obligatory course Business Information Systems at the Faculty of Economics and Business Zagreb (Figure 1).



Source: Authors' analysis

Fig. 1 Demographic characteristics of the respondents

B. Research Results

Table 1 shows what respondents consider an embedded system is. They could choose more than one answer, e.g. POS devices, Smartphone, Data base, ATM, Personal computers, Safe alarms, Peacemaker, SIM cards, eBooks, Traffic lights, Video games, MP3, Console for playing video games, Presentation pointer and Dishwasher machine. The highest percentage of respondents consider that POS device (48,57%), Smartphone (45,19%) and Data base (42,08%) present embedded systems. One quarter of the respondents believe that Personal computers (25,19%) and Safe alarms (24,16%) present embedded systems. The lowest percentage of respondents choose MP3, Console for playing video games, Presentation pointer and Dish machine as devices with embedded systems.

In your opinion, what present embedded system?	Number of respondents	%
POS devices	187	48,57%
Smartphone	174	45,19%
Data base	162	42,08%
ATM	141	36,62%
Personal computers	97	25,19%
Safe alarms	93	24,16%
Peacemaker	84	21,82%
SIM cards	84	21,82%
eBooks	64	16,62%
Traffic lights	63	16,36%
Video games	59	15,32%
MP3	49	12,73%
Console for playing video games	46	11,95%
Presentation pointer	46	11,95%
Dish machine	35	9,09%

Source: Authors' analysis

Table 1 Devices that present embedded systems

Table 2 presents different usage of GPS technology for the city sightseeing. The respondents could choose more than one answer while answering the questionnaire, e.g. Defining route between two addresses in the city; Discovering city sights close to your location; Monitoring and forecasting dense traffic and View of all restaurants on selected city area. Most of the respondents (83,12%) stated that defining route between two addresses in the city is one of the most important usage of the GPS technology for the city sightseeing. Half of the respondents excerpt that usage of the GPS technology is important in discovering city sights close to your location (47,01%). The lowest percentage of respondents named that usage of GPS technology can be useful for monitoring and forecasting dense traffic (23,12%) and finding restaurants on selected city (21,82%).

In table 3 Possession and usage of different devices is described. Respondents could choose more than one offered answer, e.g. Smartphone, Laptop, WIFI device, MP3, GPRS, Tablet, Playing console, SmartTV, Cyclocomputer and eReader. Almost all respondents (94,03%) have smartphone.

Usage of GPS technology	Number of respondents	%
Defining route between two addresses in the city	320	83,12%
Discovering city sights close to your location	181	47,01%
Monitoring and forecasting dense traffic	89	23,12%
View of all restaurants on selected city area	84	21,82%

Source: Authors' analysis

Table 2 Usage of GPS technology

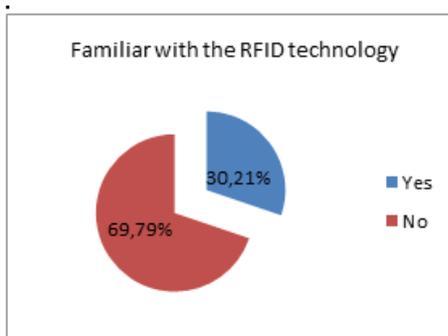
Possession and usage of different devices	Number of respondents	%
Smartphone	362	94,03%
Laptop	304	78,96%
WIFI device	278	72,21%
MP3	202	52,47%
GPRS	90	23,38%
Tablet	79	20,52%
Playing console	74	19,22%
SmartTV	54	14,03%
Computer for the bicycle	24	6,23%
eReader	15	3,90%
Nothing	6	1,56%

Source: Authors' analysis

Table 3 Possession and usage of different devices

Laptop (78,96%) and WIFI device (72,21%) are also used by high percentage of students. Half of respondents have MP3 device (52,47%). GPRS, Tablet and Playing console are devices that around 20% of respondents possess and use. Cyclocomputers (6,23%) and eReaders (3,90%) are really rarely used.

Figure 2 shows results about students' knowledge regarding RFID technology. Most of them stated that they are not familiar with RFID technology (69,79%). Only 30,21% respondents know what is RFID technology.



Source: Authors' analysis

Fig. 2 Respondents' knowledge about RFID technology

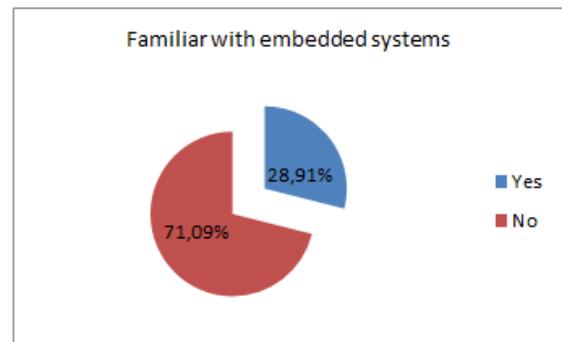
Although, respondents do not know what RFID technology is, most of them estimate that RFID is mostly used in Logistics and transport industry (66,23%) (Table 4). Just 18,44% of respondents consider that RFID is widely used in Tourism industry. The lowest percentage of respondents believe that RFID technology is primarily used in Construction industry (5,71) and in Textile industry (2,34%).

Industries	Number of respondents	%
Logistics and transport industry	255	66,23%
Tourism industry	71	18,44%
Food industry	27	7,01%
Construction industry	22	5,71%
Textile industry	9	2,34%
TOTAL	384	100%

Source: Authors' analysis

Table 4 Usage of RFID in different industries

Figure 3 shows results about students' knowledge regarding embedded systems. Most of them stated that they are not familiar with embedded systems (71,09%). Only 28,91% respondents know what embedded systems are.



Source: Authors' analysis

Fig. 3 Respondents' knowledge about embedded systems

In Table 5, data about usage of embedded systems by individuals is presented. The respondents could choose more than one offered answer, e.g. Digital thermometer, Parking sensors, eReader, Blood Glucose Monitoring Device, Peacemaker, 3D printer and Robot vacuum cleaner. The highest percentage of respondents use digital thermometer (60,00%) and parking sensors (47,79). Quarter of all respondents use also eReader (27,53%) and Blood Glucose Monitoring Device (23,38%). The lowest percentage of respondents indicates that they use 3D printer (11,95%) and Robot vacuum cleaner (11,69%).

Usage of other embedded systems	Number of respondents	%
Digital thermometer	231	60,00%
Parking sensors	184	47,79%
eReader	106	27,53%
Blood Glucose Monitoring Device	90	23,38%
Peacemaker	62	16,10%
3D printer	46	11,95%
Robot vacuum cleaner	45	11,69%

Source: Authors' analysis

Table 5 Usage of embedded systems

V. DISCUSSION

A. Estimating awareness and understanding of pervasive computing technology

The conducted research was aimed at young highly educated demographic so when interpreting the presented results we

should take into account that the targeted demographics had better access to information about the current state of the art in economics, business and partially to IT and consumer electronics. From table 3 we can also conclude that this targeted demographic is well involved in the trends in electronics since 94,03% of them own and use smartphones. It is safe to presume that due to these characteristics the targeted segment of users are in better position than most of the other demographics that have somewhat lesser access to information through academic channels. This is why we presumed that the results will show adequate understanding and knowledge about technology, and at least some knowledge about various technologies used for pervasive computing.

This is the reason why the first part of the survey resembled an examination test where some of the possible answers were correct and others incorrect. This is why in Table 1 there are correct and incorrect options to choose an answer from to the question what is considered an embedded system. The correct devices that contain or present themselves as embedded systems are POS devices, smartphones, ATMs, safe alarms, pacemakers, SIM cards, traffic lights, video games consoles, presenter pointers and dishwashers. Unfortunately, none of these devices was recognized by majority of respondents as embedded systems. Best scores for POS devices and smartphones are below 50%. It is encouraging, though, that most of the incorrect answers were recognized as incorrect in much greater extent, except from databases that had the highest score of 25% as misidentified embedded system. This means term embedded systems is not overly familiar to the surveyed demographics as they are not able to recognize its real world implementations. Further evidence for this is given in Figure 3 where less than a third of responses claim that they are familiar with embedded systems. Similar results are found for presented RFID technology shown in Figure 2.

If we take a look at more factographic questions that are being taught at various courses during their studies, results are much better as expected. An example is the question of where RFID technology is mostly used with results shown in table 4. This is a piece of information taken from the obligatory undergraduate course Business information systems that all of the respondents took part in that same year. Two thirds of respondents have recognized the most typical industry RFID technology is used in. Usage of GPS technology yielded similar results, where over 80% of students recognized one typical applications of this technology, while other correct uses (discovering city sights in vicinity and traffic density maps) were less recognized. Finally viewing static data on a map that does not require GPS technology was recognized as non GPS application by similar number of students.

Overall, we may conclude that students have some overall declarative knowledge about the mentioned technologies but do not have enough understanding to recognize their usage in real world applications, which may also indicate the lack of awareness of their availability, advantages and disadvantages and potentials.

B. Usage of products and services containing embedded systems, Internet and mobile technologies

Finally, in order to estimate the level of usage of products and services that rely on pervasive computing technologies and principles we focused on what devices students possess and use (shown in Table 3) and what other devices they use or know how to operate (shown in Table 5).

If we take a look at the devices and consumer electronics that respondents own we can see that all of the market segments discussed in Section III are represented. This means that there is a good foundation for development of additional services related to any of the customer electronic devices present on the market which is one of the bases of pervasive computing. Even lesser used devices with very particular purposes are represented as we can see in Table 5. Most important is the use of smartphones and laptops with over 90% and almost 80% of respondents use them, respectively.

Since smartphones present an intermediary form between embedded systems and general-purpose devices [26] it is safe to presume that potential for further use of pervasive computing is significant. Earlier studies [27, 28] have shown that student use various mobile applications and innovative mobile services. This is why better understanding of underlying technologies is called for, or in the case of pervasive computing, implementations that will minimize the risks of technologies used thus increasing further the use of applications and services.

VI. CONCLUSIONS

Pervasive computing is a computing paradigm where most of devices have some computing capability and interconnectivity which results in creation of proactive and self-tuning environments and devices that can augment personal knowledge and decision making abilities of its users, while requiring as little direct user interaction as possible. In the recent years various examples of pervasive computing implementations have been created which is partly a result of technological innovation cycle which was initiated during the latest global economics crisis in 2008. All of the related markets have shown high level of resilience to economic fluctuations, both customer electronics markets, software markets etc.

In this paper we investigated this phenomenon from the customers' standpoint trying to find the explanation for the positive trends in related industries and the incentive that customers through market absorption supported the process of innovation and development of various segments of pervasive computing.

The goal of this paper was twofold. Firstly, by using the survey targeted at highly educated young demographics investigate the level of technological understanding, and secondly to investigate practical involvement of this demographic in increasing trends of using products and services related to embedded systems, mobile technologies and Internet technologies.

The results have shown that there is insufficient level of in-depth understanding of used technologies, but this lack of awareness does not diminish the level of practical involvement and interest for products and services these technologies offer. This is in accordance with one of guidelines of pervasive computing where providing innovative and convenient new functionality does not require the user to understand the inner working. This opens a number of challenges concerning security, privacy, trustworthiness and disaster recovery issues. The incentive to deal with these issues is on the pervasive computing solutions.

Still, correlation between understanding of underlying technologies and rate of acceptance of products and services it enables remain for future investigation. Also, understanding the limitations of customer acceptance needs to be further explored.

REFERENCES

- [1] J.H. Abawajy, (2009), "Advances in pervasive computing", *International Journal of Pervasive Computing and Communications*, Vol. 5 Iss 1 pp. 4 – 8.
- [2] P. Abrahamsson, A.Hanhineva, H. Hulkko, T. Ihme, J. Jääliñoja, M. Korkala, J.Koskela, P. Kyllönen, O. Salo (2004) "Mobile-D: an agile approach for mobile application development", In Proceeding Companion to the 19th annual ACM SIGPLAN conference on Object-oriented programming systems, languages, and applications (OOPSLA '04), ACM New York, NY, USA, pp. 174-175.
- [3] K. Cousins and D. Robey, (2015). "Managing Work-Life Boundaries with Mobile Technologies: An Interpretive Study of Mobile Work Practices". *Information Technology and People*, 28(1), pp. 34-71.
- [4] A. del Aguila-Obra, and A. Padilla-Melendez (2006). "Organizational Factors Affecting Internet Technology Adoption". *Internet Research*, 16(1), pp. 94-110.
- [5] M. De Saullés, and D. S., Horner, (2011). "The Portable Panopticon: Morality and Mobile Technologies". *Journal of Information, Communication and Ethics in Society*, 9(3), pp. 206-216.
- [6] S. Dutta, and B. Bilbao-Osorio(eds.) (2012). The Global Information Technology Report-Living in a Hyperconnected World. Geneva: World Economic Forum. Available: http://www3.weforum.org/docs/Global_IT_Report_2012.pdf [19/05/2015]
- [7] C. Ebert and C. Jones (2009). "Embedded Systems: Facts, Figures and Future". *Computer*. Issue April 2009, IEEE Computer Society, pp. 42–52.
- [8] M. Fleck, M. Frid, T. Kundberg, E. O'Brien-Strain, R. Rajani and M. Spasojevic (2002), "From informing to remembering: ubiquitous systems in interactive museums", *IEEE Pervasive Computing*, Vol. 1 No. 2, pp. 13-21.
- [9] T. Giuli, D. Watson and K. V. Prasad (2006), "The last inch at 70 miles per hour", *IEEE Pervasive Computing*, Vol. 5 No. 4, pp. 20-7.
- [10] H. Hagras, V. Callaghan, M. Colley, G. Clarke, A. Pounds-Cornish and H. Duman (2004), "Creating an ambient-intelligence environment using embedded agents", *IEEE Intelligent Systems*, Vol. 19 No. 6, pp. 12-20.
- [11] T. Hansen, J. Bardram and M. Soegaard, (2006), "Moving out of the lab: deploying pervasive technologies in a hospital", *IEEE Pervasive Computing*, Vol. 5 No. 3, pp. 24-31.
- [12] K. Harle and A. Hopper (2005), "Deploying and evaluating a location-aware system", Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services, Seattle, WA, pp. 219-32.
- [13] A. Helal, W. Mann, H. Elzabedani, J. King, Y. Kaddourah, and E. Jansen (2005), "Gator tech smart house: a programmable pervasive space", *IEEE Computer Magazine*, March, pp. 66-74.
- [14] A. Holmes, H. Duman, and A. Pounds-Cornish (2002), "The iDorm: gateway to heterogeneous networking environments", in Proceedings of International Test and Evaluation Association (ITEA) Workshop Virtual Home Environments, ITEA Press, Fort Berthold, ND, pp. 30-7.
- [15] Infinity Research Limited (2014) "Global Consumer Electronics Market 2014-2018". March 2014.
- [16] Internet World Stats. Available: <http://www.internetworldstats.com/stats.htm> [25/05/2015]
- [17] R. Jain and J. Wullert (2002), "Challenges: environmental design for pervasive computing systems", Proceedings of the 8th Annual International Conference on Mobile Computing and Networking, Atlanta, GA, pp. 263-70.
- [18] E. H. Marinho and R. F. Resende (2012). "Quality factors in development best practices for mobile applications", in Proceedings of the 12th international conference on Computational Science and Its Applications (ICCSA'12)- Volume Part IV, Springer-Verlag Berlin, Heidelberg, pp. 632-645.
- [19] R. McIvor, M. McHugh, M. and C. Cadden (2002). "Internet Technologies: Supporting Transparency in the Public Sector". *International Journal of Public Sector Management*, 15(3), pp. 170-187.
- [20] D. Patterson (2005), "20th century vs 21st century C&C: the SPUR manifesto", *Communication of the ACM*, Vol. 48 No. 3, pp. 15-16.
- [21] M. Schraefel, G. Hughes, H. Mills, G. Smith, T. Payne and J. Frey (2004), "Breaking the book: translating the chemistry lab book into a pervasive computing lab environment", In Proceedings of the 2004 Conference on Human Factors in Computing Systems, Vienna, pp. 25-32.
- [22] D. Singh Negi, (2014). "Using Mobile Technologies in Libraries and Information Centers". *Library Hi Tech News*, 31(1), pp. 14-16.
- [23] Truffle Capital, Top 100 European Software vendors: the best software companies, available at <http://www.truffle100.com/2014/countries.php>, 2015.
- [24] Vison Mobile (2013) "The European App Economy 2013". Spetember 2013.
- [25] F. Zambonelli and M. Viroli, (2011), "A survey on nature-inspired metaphors for pervasive service ecosystems", *International Journal of Pervasive Computing and Communications*, Vol. 7 Iss 3 pp. 186 – 204.
- [26] J. L. Gómez-Barroso, M. Bacigalupo, S. G. Nikolov, R. Compañó, C. Feijóo, (2012) "Factors required for mobile search going mainstream", *Online Information Review*, Vol. 36 Iss: 6, pp.846 – 857.
- [27] Z. Pozgaj, N. Vlahovic, (2011): Students' readiness for informal learning using Web 2.0 services; In: Cicin-Sain, M., Sunde, J. (2011): *International Journal of Knowledge and Learning*, 2011, vol.7, no.1-2, pp. 113-29.
- [28] N. Vlahovic, J. Zoroja, V. Bosilj Vuksic (2013) Research on customers' attitudes towards smartphones as an intermediate between embedded systems and general purpose devices; In: Merkač Skok, M. & Cingula M. (eds) Knowledge and business challenge of globalisation in 2013: conference proceedings of the 5th international scientific conference, Faculty of Commercial and Business Sciences, Celje, 14th-15th November 2013, pp. 218 – 224.

Exploiting the Interpretability of Fuzzy Rule-Based Classifiers for Analyzing Hyperspectral Remotely Sensed Data

Dimitris G. Stavrakoudis, Stelios K. Mylonas, Charalampos A. Topaloglou, John B. Theocharis, and Paris A. Mastorocostas

Abstract—This paper showcases the use of fuzzy rule-based classification systems (FRBCSs) for analyzing hyperspectral data in the context of remote sensing classification tasks. First, the wavelet packet decomposition (WPD) is applied to the original data, in order to obtain higher-order features that describe the spectral changes within specific ranges of the electromagnetic spectrum. Subsequently, an advanced genetic FRBCS (GFRBCS) of the literature is applied, namely, the Fast Iterative Rule-based Linguistic Classifier (FaRLiC), which is able to produce high-performing fuzzy rule bases with very low structural complexity. Ultimately, the information provided by FaRLiC is exploited in order to discover useful hidden relations between the input features, which enable the easy discrimination between specific classes. As such, the employed model offers new tools for specialized photointerpretation, as well as the possibility for devising customized indices in the future. The procedure is employed here considering a Hyperion satellite image over a forested area in northern Greece, primarily focusing on the discrimination between two pine species.

Keywords—Genetic fuzzy rule-based classification systems (GFRBCS), hyperspectral imagery, interpretable classification models, wavelet packet decomposition.

I. INTRODUCTION

HYPERSPECTRAL spectroscopy constitutes a unique tool for observing the spectral properties of objects. Hyperspectral sensors collect several (typically 200 or more) narrow spectral bands, from the visible to the short-wave infrared portions of the electromagnetic spectrum, providing an almost continuous spectral reflectance signature. Contrary

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) – Research Funding Program “ARCHIMEDES III: Investing in knowledge society through the European Social Fund”.

D. G. Stavrakoudis is with the School of Forestry and Natural Environment, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece (phone: +30-2310-992689; fax: +30-2310-992677; e-mail: jstavrak@auth.gr).

S. K. Mylonas, C. A. Topaloglou, and J. B. Theocharis are with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece (e-mails: smylonas@auth.gr, chartopal@gmail.com, theochar@eng.auth.gr).

P. A. Mastorocostas is with the Department of Computer Engineering, Technological Educational Institute of Central Macedonia, Serres 62124, Greece (e-mail: mast@teiser.gr).

to multispectral data, hyperspectral data have been proven capable of producing both genus-level and species-level classifications [1]. Particularly for land cover classification of forests—where typically different species of the same genus coexist—it has been shown that hyperspectral imagery can significantly increase the classification accuracy [2]. Accordingly, it has been observed that the spectral profiles of plants are indicative of various vegetation properties, a fact that has led the respective research community in devising various specific hyperspectral indices [3], [4].

In the context of land cover classification, the bands of a hyperspectral image are typically considered as separate features. Since however successive hyperspectral bands are highly correlated, it would make sense to also observe the changes of reflectance (or radiance) within smaller ranges of the full spectral space. Such an analytical framework is provided by the wavelet transform, which employs a multi-resolution analysis of a signal in the time-frequency domain [5], which in our case is rather the spectral-frequency domain. The wavelet transform has lately proven to be a useful tool for increasing classification accuracy using hyperspectral imagery [6], [7]. In this paper we employ a more general form of the wavelet transformation, namely, the wavelet packet decomposition (WPD).

The objective of this paper is to discover useful relations between the input features (either the original bands of the image or the WPD-derived features) that can discriminate specific classes of the problem. This is accomplished considering the so-called FaRLiC algorithm [8], which creates classification models comprising a small set of fuzzy IF-THEN rules that resemble the form of the categorical propositions made by humans. These rules are subsequently analyzed in order to infer the desired relationships in the feature space.

The rest of the paper is organized as follows. Section II describes the study area and the dataset formulation that was used for the analysis. Section III offers a short description of the WPD, whereas Section IV briefly describes the employed FRBCS. The experimental results are reported in Section V and the paper concludes in Section VI, with some final remarks.

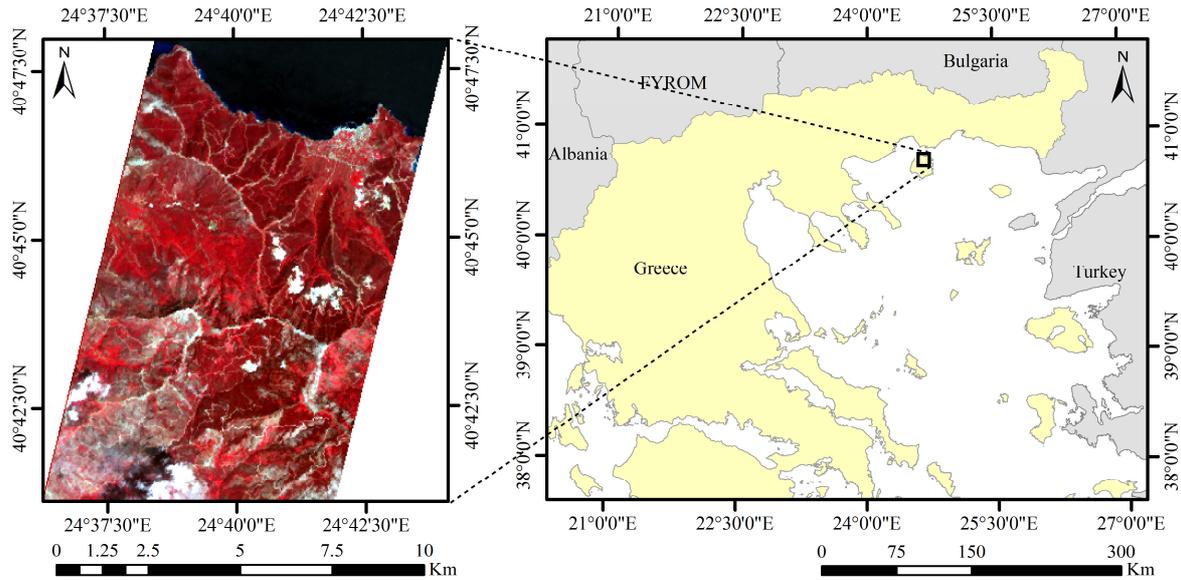


Fig. 1 The study area, along with a false-color composite of the Hyperion image.

II. STUDY AREA AND DATASET FORMULATION

The study area is the northern part of Thasos, Greece's most northerly island (Fig. 1). *Pinus brutia* is the dominant tree species at lower elevations (0 to 800 m), whereas *Pinus nigra* is found at higher altitudes. On August 1, 2003, a Hyperion image (level 1 radiometric product) covering a part of the island from north to south was acquired. The image exhibits 30 m spatial resolution and comprises a total of 242 bands in the visible to shortwave infrared portion of the electromagnetic spectrum (400–2500 nm). The dataset is formed removing the uncalibrated bands (bands with no data), thus keeping only the 196 useful bands. A false-color composite of the Hyperion image inside the boundaries of study area is depicted in Fig. 1.

After extensive field survey and careful photointerpretation, 1000 points (pixels) were identified and labeled into 6 classes: *Pinus brutia*, *Pinus nigra*, deciduous trees, other vegetation, non-vegetated areas, and water. The last class was visually identified in the image and was included in the classification scheme for the sake of completeness in the resulting thematic map. These labeled pixels formulate the reference dataset, which is used for creating the classifier and testing its accuracy.

III. WAVELET PACKET DECOMPOSITION

The wavelet transform is based on the concept of multi-resolution analysis of a signal [5]. The transformation filters the original signal with translations and dilations of a basic function called *mother wavelet* $\psi(t)$. The wavelet theory is traditionally presented considering time-varying signals and hence uses the time variable notation t . In our case, however, the signal is defined in the spectral range and therefore all signals and functions will be represented using the wavelength variable λ in the following.

The mother wavelet $\psi(\lambda)$ is used in order to form a wavelet basis that will represent the signal, through translations (displacements) and scaling of the mother wavelet. Typically binary partitions of the space are considered, in which case the so-called dyadic wavelets are defined recursively as:

$$\psi_{j,n}(\lambda) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{\lambda - 2^j n}{2^j}\right), \quad (1)$$

where 2^j is the scaling factor and n the translational factor. The functions $\psi_{j,n}(\lambda)$ constitute the orthonormal bases in a subspace W_j , which describes the high frequencies of the signal at the decomposition level j . The signal is decomposed considering an equivalent low frequency subspace V_j (which is orthogonal to W_j), using a corresponding set of wavelet bases $\phi_{j,n}(\lambda)$. The *wavelet coefficients* are defined as the inner product of the signal with the respective bases and can be used directly to represent the signal in the time-frequency domain.

For discrete signals (like those provided by hyperspectral sensors) we typically employ the discrete wavelet transform (DWT). The DWT initially decomposes the original signal into the two orthogonal spaces V_1 and W_1 . The low frequency space V_1 is subsequently decomposed into two respective spaces V_2 and W_2 and the process is repeated for a number of predefined levels, thus forming a wavelet tree. In each level, the decomposition is performed through a convolution with the discrete low-pass and high-pass filters h and g , respectively, and the signal is down-sampled with a factor of 2. The filters h and g form the so-called analysis filter bank and their coefficients are determined by the wavelet functions $\phi_{j,n}(\lambda)$ and $\psi_{j,n}(\lambda)$, respectively.

In this paper we consider the generalization of the DWT, namely, the wavelet packet decomposition (WPD). The WPD is similar to DWT, with the difference that both subspaces V_j and W_j are decomposed into two subspaces in the next level. As such, it provides more information at different resolutions,

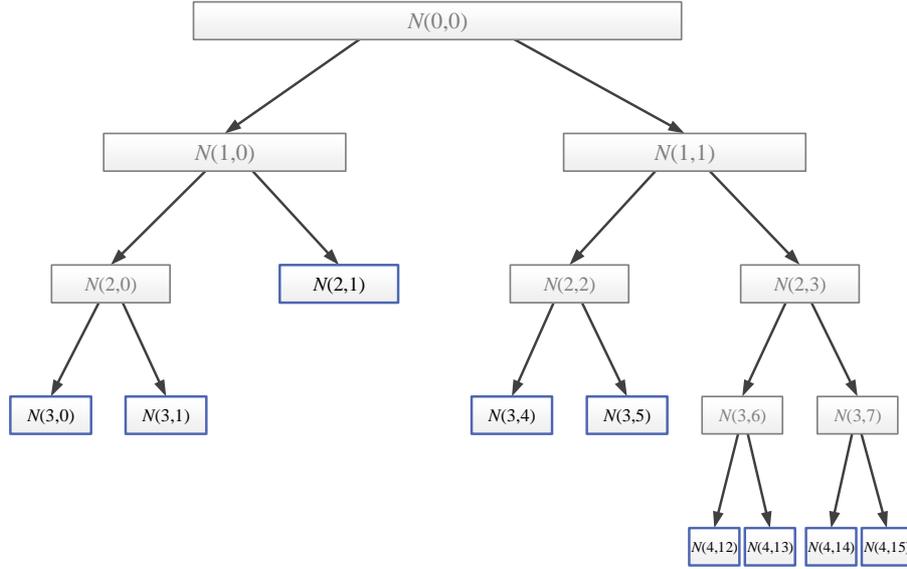


Fig. 2 The optimal WPD produced by WPD-FuzCoC. The selected nodes that form the valid WPD are shown in blue.

allowing to observe increasingly higher frequency signal changes. Traditionally, only the wavelet packet coefficients (WPCs) of the highest level are considered. However, we used the so-called WPD-FuzCoC algorithm [9], which allows the selection of WPCs from different levels, providing an optimal but valid WPD. We considered four levels of decomposition, using a reverse biorthogonal wavelet basis (rbio3.7). The latter has been chosen because it resulted to the highest classification accuracy, although all bases exhibited similar performance. The optimal WPD that was derived through this process is shown in Fig. 2, where the selected nodes are highlighted with blue color. The notation $N(j,k)$ is used to represent the k th node ($k = 0, \dots, 2^j - 1$) of the level j ($j = 1, \dots, 4$) in the decomposition tree.

Conceptually, the derived WPCs can be thought as the description of the original hyperspectral signals (defined in the vector space with 196 components in our case) in the spectral-frequency domain. For example, the node $N(1,0)$ in Fig. 2 comprises 98 WPCs, each representing the low-frequency changes every two consecutive bands, whereas the coefficients of node $N(1,1)$ describe the high-frequency changes in the same ranges. The nodes of the second level describe the changes every 4 bands, the nodes of the third level every 8 bands and so on. In the following, each coefficient is represented using the notation $WPC(j,k,c)$ [spectral range]. For example, the notation $WPC(3,1,13)$ [1376.55, 1447.14] represents the 13th WPC of the second node in the third level $N(3,1)$, which describes the high-frequency changes (because it is the second node with respect to the previous level) in the spectral range [1376.55, 1447.14] (because the third level is defined in ranges of 8 consecutive bands).

Through the valid WPD of Fig. 2, 196 WPCs have been extracted. The full feature space is created considering the original bands of image as well, thus comprising 392 dimensions in total.

IV. CLASSIFICATION MODEL

For classification purposes we used a fuzzy rule-based classifier called FaIRLiC [8]. FaIRLiC is a GFRBCS characterized by its ability to produce fuzzy classification rule bases with very low complexity, even for very high-dimensional feature spaces comprising hundreds or even thousands of features. This section briefly describes FaIRLiC's classification model, whereas a thorough description of its learning algorithm can be found in the aforementioned reference.

We consider a classification problem defined in an N -dimensional feature space $\mathbf{X} = [X_1, \dots, X_N] \in \mathfrak{R}^N$ and an M -dimensional class space $\mathcal{C} = \{C_1, \dots, C_M\}$. FaIRLiC comprises a number of fuzzy IF-THEN rules, each defining a fuzzy relation between the feature (antecedent part – IF) and class (consequent part – THEN) spaces. Initially, each input variable (feature) is partitioned into ℓ fuzzy sets $\{L_1^i, \dots, L_\ell^i\}$ with equivalent linguistic labels (for example, small, high, etc.). A typical approach is to consider for each input variable a uniformly distributed partition of fuzzy sets with triangular membership functions, as shown in Fig. 3. A rule relates fuzzy expressions on the feature space with the class space:

$$R^k: \text{IF } X_1 \text{ is } A_1^k \text{ AND } \dots \text{ AND } X_N \text{ is } A_N^k \\ \text{THEN class } C^k \in \mathcal{C} \text{ with } r^k \quad (2)$$

where each input variable X_i takes as a value a fuzzy set A_i^k and the rule also includes a confidence degree r^k in its consequent, representing the confidence of the classification in the class label of the consequent (C^k) and calculated using a training set of labeled patterns. FaIRLiC assumes linguistic fuzzy rules in the so-called disjunctive normal form (DNF). This means that each fuzzy set A_i^k in the antecedent part of the rule is not necessarily only one of the ℓ predefined fuzzy sets,

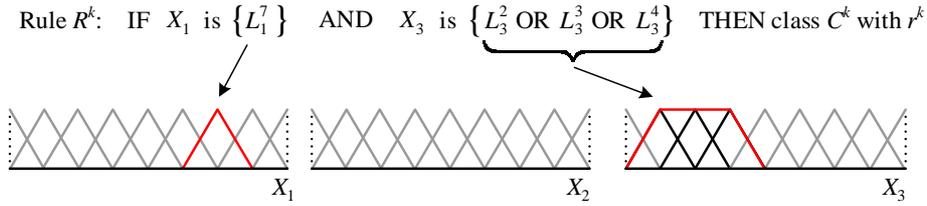


Fig. 3 An example of a fuzzy rule for a hypothetical classification problem with three features.

but it may also be a composite fuzzy set joining multiple consequent terms with the OR disjunctive operator, for example $A_i^k = \{L_i^2 \text{ OR } L_i^3 \text{ OR } L_i^4\}$. Moreover, some input variables in a rule are allowed to be absent (or inactive), which in fuzzy terms has the meaning of a “don’t care” fuzzy set (a fuzzy set with a unity membership grade for its entire universe of discourse).

An example of a fuzzy rule for a hypothetical classification problem with three features is depicted in Fig. 3. In this example, each feature has been uniformly partitioned into 9 triangular fuzzy sets, with the two outermost sets having half the width than the others. The antecedent part for the rule is formed considering only two of the three features, whereas the second one (X_2) is inactive. The first feature takes as value a single linguistic term L_1^5 , whereas for the third feature a composite fuzzy set is considered, joining three consecutive terms with the OR fuzzy operator.

The whole fuzzy rule base comprises a number of rules in the form of (2), where multiple rules describing a single class are allowed to coexist. The fuzzy reasoning mechanism determines how the classification of an input vector $\mathbf{x} \in \mathbf{X}$ is derived. For each rule, all fuzzy sets participating in its antecedent part provide a degree $\mu_{L_i^k}(x_i)$ in $[0,1]$, as defined by their membership functions (triangular in our case). The firing of the k th rule (also known as matching degree) is calculated through:

$$\mu^k(\mathbf{x}) = \bigcap_{i=1}^N \left\{ \bigcup_{q=1}^{\ell^k} \mu_{L_i^q}(x_i) \right\} \quad (3)$$

where \cap and \cup denote the AND and OR operators, respectively. In effect, the membership degree of each (possibly composite) fuzzy set is calculated applying the OR operator and, subsequently, the total rule firing is determined employing the AND operator between all active features. In FaIRLiC, the AND operator is implemented through the minimum operator, whereas the OR operator is implemented through the bounded sum, defined for two membership values a and b as:

$$\text{bs}(a,b) = \min(1, a+b) \quad (4)$$

With this choice, the composite fuzzy set is mathematically equivalent to the circumscribed trapezoidal fuzzy set, as shown in the example of Fig. 3 for the third feature. The input vector is ultimately assigned taking into consideration all K rules of the rule base, through the so-called maximum voting scheme:

$$\mathbf{x} \rightarrow C_j \in \mathcal{C}, \quad j = \arg \max_{m=1, \dots, M} \sum_{R^k | C^k = C_m} \mu^k(\mathbf{x}) \cdot r^k, \quad (5)$$

that is, to the class achieving the highest cumulative activation degree among the rules. Note that (5) also considers the certainty degrees r^k of each rule, which is multiplied with its firing. To this end, the certainty degrees can be viewed as weights, representing the relative importance of each rule in the classification decision.

The whole rule base is constructed through an iterative process, with a single rule being extracted in each iteration. The whole process is coordinated by a pattern weighting scheme, which guides new rules in exploring the regions of the feature space not covered by previously extracted rules. The rule extraction algorithm extensively evaluates a feature selection-based metric, which eventually leads to the extraction of fuzzy rules with few active features, even for very high-dimensional classification tasks. Finally, a genetic tuning post-processing stage fine-tunes the definitions of the triangular membership functions (without changing their relative order), in order to further increase the classification accuracy of the derived model.

V. EXPERIMENTAL RESULTS

The objective of this paper is to showcase how the information provided by an interpretable classification model can be used for understanding the discriminating relations between the input features. Before that, however, we briefly present the numerical results obtained by FaIRLiC in comparison to other classifiers, in order to prove its validity in solving the classification task considered.

FaIRLiC is compared here with three other GFRBCSs (FeSLiC [10], 2SLAVE-2 [11], and SGERD [12]), three crisp rule-based classifiers (C4.5 Rules [13], RIPPER [14], and SLIPPER [15]), the C4.5 decision tree [13], and the support vector machine (SVM) classifier [16], which has been shown to exhibit very high classification accuracies in a number of classification tasks. In order to derive unbiased conclusions, we used a 5-fold partitioning of our dataset. For classifiers whose learning algorithm has stochastic characteristics (all but C4.5 and SVM), we performed six independent runs on each partition, using different random seeds. In each case, the average results over these 30 runs are given. For FaIRLiC and 2SLAVE-2, each input variable has been partitioned into 9 uniformly distributed fuzzy sets.

TABLE I. AVERAGE RESULTS OBTAINED FOR ALL CLASSIFIERS CONSIDERED (CLASSIFICATION ACCURACY ON THE TESTING SET AND STRUCTURAL CHARACTERISTICS). THE SECOND COLUMN ALSO REPORTS THE AVERAGE TESTING ACCURACY CONSIDERING ONLY THE ORIGINAL BANDS OF THE IMAGE.

Classifier	Bands Only		Full feature space: Bands + WPC		
	Testing accuracy (%)	Testing accuracy (%)	#R	#GUF	#F/R
FaIRLiC	84.85	88.20	6.97	19.10	2.82
FeSLiC	84.63	85.52	9.13	29.27	3.34
SGERD	78.12	81.02	7.90	13.40	2.00
2SLAVE-2	80.92	85.37	11.60	86.97	8.70
C4.5	84.20	85.10	34.60	28.20	7.78
C4.5 Rules	83.40	85.20	15.00	23.60	4.04
RIPPER	82.10	85.90	23.40	34.00	2.10
SLIPPER	85.50	88.10	241.40	142.60	1.54
SVM	85.90	90.80	–	–	–

#R = number of rules; #GUF = number of globally used features; #F/R = average number of features per rule.

Table I presents the obtained results, averaged over the different runs (5 for C4.5 and SVM, 30 for all other classifiers). The table reports the testing accuracy, along with the most important structural characteristics of the obtained models: number of rules (#R), average number of features per rule (#F/R), and globally used features (#GUF). The latter describes the number of features selected by at least one rule in the rule base. The second column of the table also reports the each classifier's average testing accuracy when considering only the bands of the image, without the WPCs.

A first remark is that the accuracy of all classifiers increases with the addition of the WPC features, despite the fact that the feature space is effectively doubled. The increase in most cases is substantial, which proves that the extra features derived from the WPD are indeed informative. FaIRLiC achieves the second highest accuracy, being inferior to only the SVM classifier, the model of which however cannot provide any linguistic interpretation. Comparatively to all other interpretable classifiers, FaIRLiC and SGERD produced the simplest models. Nevertheless, the latter constructs over-simplistic models, exhibiting by far the worst classification accuracy amongst all classifiers tested. The comparison of Table I proves that FaIRLiC can indeed produce very simple and accurate fuzzy rule bases.

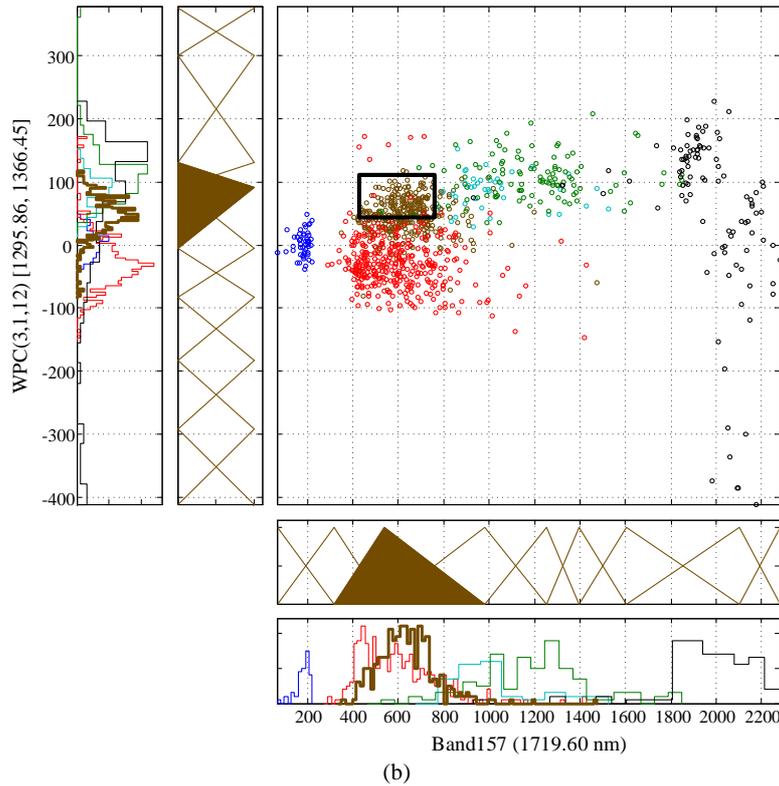
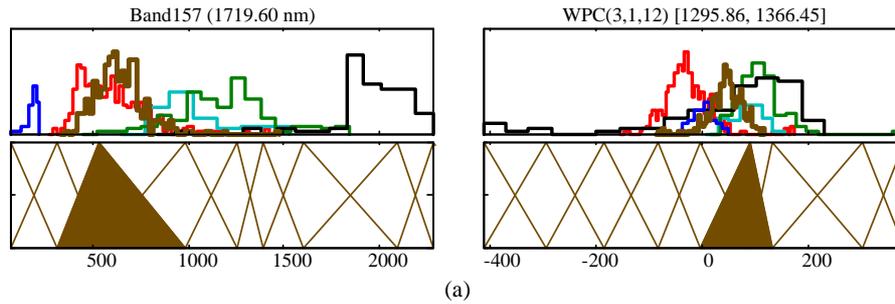
We now concentrate on the FaIRLiC model that exhibits the highest classification accuracy, which comprises six fuzzy classification rules (one for each class) and used globally 11 features (7 bands and 4 WPCs). The major difficulty of the specific classification task is the discrimination between the two pine species of the area (*Pinus brutia* and *Pinus nigra*). As an example, the analysis here is presented based on the rule created for the *Pinus brutia* class.

Fig. 4(a) provides a visual representation of the rule's premise part, with the first line of subplots depicting the histograms of the classes' patterns in the reference set. The fuzzy partitions are not uniformly distributed because of the tuning stage mentioned in the previous section, which fine-tunes the membership functions' definitions. The rule considers only two features the band with central wavelength 1719.60 nm and the wavelet coefficient WPC(3,1,12)

[1259.86, 1366.45]. According to the description of Section III, this coefficient quantifies the high frequency changes (second node) inside the 8 bands (third level) spectral range [1259.86, 1366.45] (nm). Fig. 4(b) depicts the respective scatter plot in the 2-dimensional space, along with the location of the 0.5-cut of the rule (that is, all points of the space for which the rule is activated with a degree of 0.5). It becomes apparent that those two features can discriminate the *Pinus brutia* class from all other classes to a high degree. It is worth noting that the WPC feature is defined in within a spectral region characterized by strong water absorption. In other words, this feature quantifies the high frequency changes of the reflectance response near a point where water absorbs all solar energy.

Linguistically, the rule of Fig. 4(a) can be described by the following phrase: "If a pixel exhibits relatively small intensity in band 157 (central wavelength 1719.60 nm) and medium to high value in feature WPC(3,1,12) [1259.86, 1366.45] (that is, medium to high frequency changes in this spectral range), then it is an area covered with *Pinus brutia*." The visual representation of the two features as grayscale images is provided in Figs. 5(b) and 5(c), respectively. For easy comparison, the thematic map is also shown in Fig. 5(a). If we observe carefully, all forest species exhibit low intensity values in band 157, with other vegetation (mainly agricultural land) and non-vegetated areas exhibiting higher intensity. At the same time, however, the *Pinus nigra* class exhibits relatively low values in the WPC feature, whereas the deciduous trees and non-vegetated areas exhibit much higher values. As such, these two features can effectively discriminate the *Pinus brutia* class from all other land cover types of the specific area.

The same conclusion can be derived if we construct a false-color composite from these two features. For example, if we use band 157 for the red channel and the WPC feature for the green channel of the color image (the blue channel is left with zero values), then we obtain the color image of Fig. 5(d). In order to increase the color contrast, the image has been enhanced by saturating the lower and upper 1% each band's histogram. This is a typical procedure employed when visualizing remotely sensed imagery. With this choice, a



Class legend:



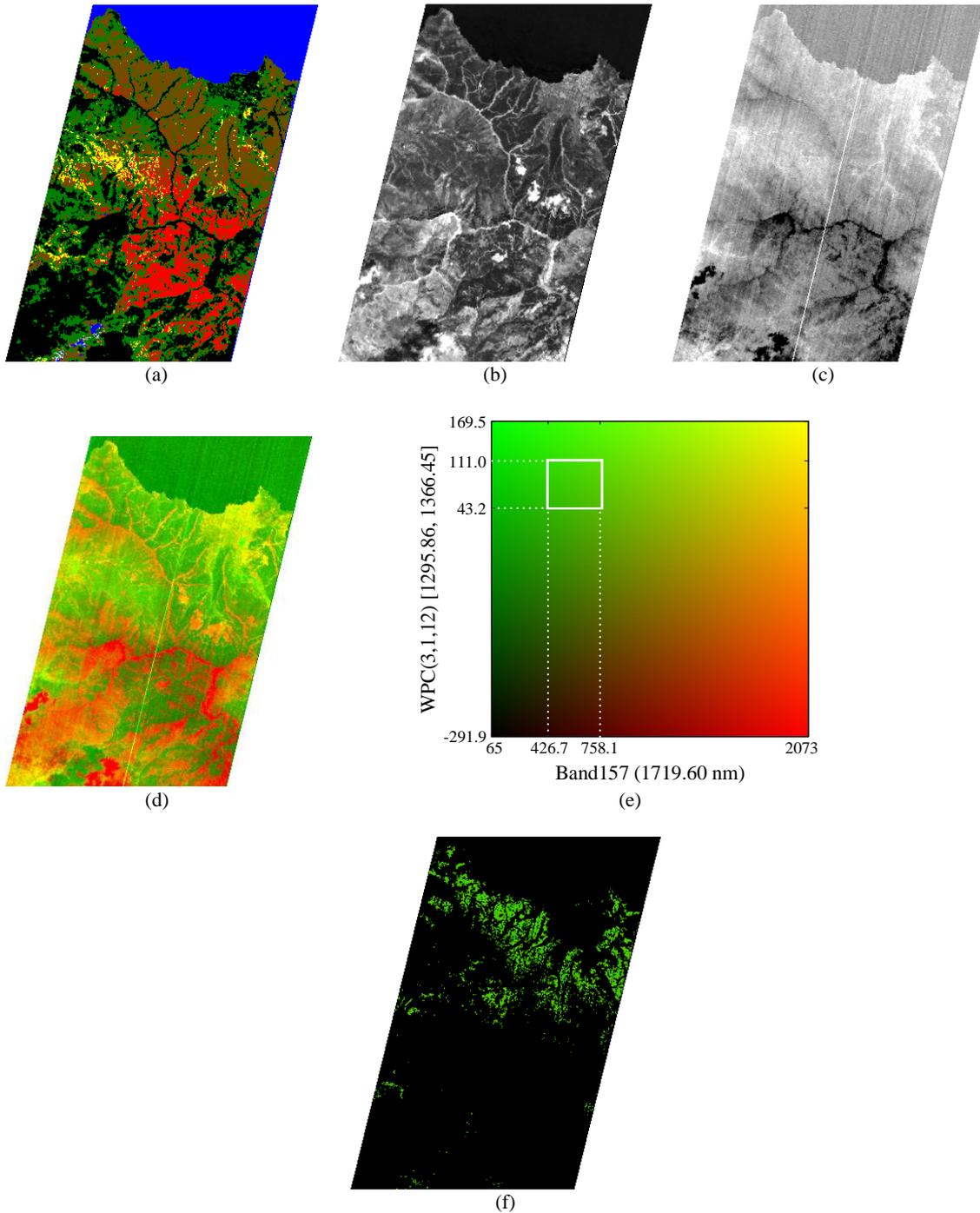
Fig. 4 The rule produced for *Pinus brutia*: (a) visual representation of the rule's antecedent part and (b) scatter plot of the reference pattern along with the rule's 0.5-cut in the feature space.

respective color map is defined in the 2-dimensional space, as shown in Fig. 5(e). The 0.5-cut of the fuzzy rule has a unique location in this color space, which is represented by the white rectangle in Fig. 5(e). This is actually the location of the black rectangle in Fig. 4(b), if we employ the image enhancement method to the specific feature space (that is, the boundaries of each feature are shrunk).

Using the aforementioned procedure, all colors in Fig. 5(d) inside the 0.5-cut represent areas covered with *Pinus brutia*. Non-vegetated areas are represented with green-yellow, orange and red colors. Deciduous trees exhibit vibrant green colors, whereas areas covered with *Pinus nigra* are represented by darker green colors. To make it more clear, Fig. 5(f) presents the same composite by zeroing the values of all pixels with colors outside the white rectangle in Fig. 5(e). Comparing

Fig. 5(f) with the thematic map of Fig. 5(a), it becomes apparent that the 0.5-cut of the fuzzy rule discriminates the class described by the rule from all others. Effectively, the color composite of Fig. 5(d) can uniquely identify the *Pinus brutia* class in the study area. Arguably, this simple rule could be generalized to other images from the same sensor, assuming that the spectral response of the sensor does not change between acquisitions.

The process presented above can be applied for the other classes of the problem, even if the respective rules comprise more than three features. As such, FaIRLiC's interpretability properties can be used to easily discriminate specific classes through photointerpretation, even for classes with high overlapping, such as species belonging to the same genus. Eventually, the interpretation provided by FRBCSs could even



Thematic map legend:



Fig. 5 Photointerpretation for the *Pinus brutia* class: (a) thematic map obtained by FaIRLiC, (b) the band of the image with central wavelength 1719.60 nm, (c) grayscale representation of the WPC(3,1,12) [1259.86, 1366.45] feature, (d) a false-color composite using the two features for the red and green channels, respectively of the image, (e) the associated color map in the 2-dimensional space, and (f) the false-color composite presented in (d), if we zero all regions outside the 5.5-cut of the fuzzy rule.

be exploited in the future for devising customized indices. In this case, however, a large number of reference spectral signatures should be collected—ideally in different areas—and

the analysis should be performed considering reflectance values (and not radiance ones as we did in this paper). In other words, FaIRLiC should be employed to hyperspectral data

obtained from carefully constructed spectral libraries. As such, the results of the analysis could be possibly generalized to different regions with similar land cover types.

VI. CONCLUSION

This paper presented the relative advantages of FRBCSs in the context of land cover classification tasks using hyperspectral imagery. In particular, the linguistic model of the employed FaIRLiC classifier revealed relations between specific input features, which enabled the discrimination of specific classes through photointerpretation. Additionally, an optimal WPD was employed in order to increase the discrimination between the classes and to observe the hyperspectral signatures within larger portions of the spectrum comprising multiple consecutive bands. Arguably, the approach followed could be possibly proven useful in the future for devising customized indices, which could easily discriminate specific classes of interest.

ACKNOWLEDGMENT

The Hyperion image was downloaded free of charge from USGS GloVis (<http://glovis.usgs.gov/>). The authors thank Prof. Ioannis Gitas for providing the reference data for the Thasos dataset.

REFERENCES

- [1] M. Govender, K. Chetty, V. Naiken, and H. Bulcock, "A comparison of satellite hyperspectral and multispectral remote sensing imagery for improved classification and mapping of vegetation," *Water SA*, vol. 34, no. 2, pp. 147–154, 2008.
- [2] D. G. Goodenough, A. Dyk, K. O. Niemann, J. S. Pearlman, H. Chen, T. Han, M. Murdoch, and C. West, "Processing Hyperion and ALI for forest classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1321–1331, Jun. 2003.
- [3] D. Haboudane, J. R. Miller, E. Pattey, P. J. Zarco-Tejada, and I. B. Strachan, "Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture," *Remote Sens. Environ.*, vol. 90, no. 3, pp. 337–352, Apr. 2004.
- [4] D. Haboudane, J. R. Miller, N. Tremblay, and P. Vigneault, "Indices-based approach for crop chlorophyll content retrieval from hyperspectral data," in *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International*, 2007, pp. 3297–3300.
- [5] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd edition. Amsterdam ; Boston: Academic Press, 2008.
- [6] L. M. Bruce and J. Li, "Wavelets for computationally efficient hyperspectral derivative analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1540–1546, Jul. 2001.
- [7] M. Z. Salvador, R. G. Resmini, and R. B. Gomez, "Detection of Sulfur Dioxide in AIRS Data With the Wavelet Packet Subspace," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 1, pp. 137–141, Jan. 2009.
- [8] D. G. Stavrakoudis, G. N. Galidaki, I. Z. Gitas, and J. B. Theocharis, "Reducing the Complexity of Genetic Fuzzy Classifiers in Highly-Dimensional Classification Problems," *Int. J. Comput. Intell. Syst.*, vol. 5, no. 2, pp. 254–275, 2012.
- [9] S. P. Moustakidis, J. B. Theocharis, and G. Giakas, "A fuzzy decision tree-based SVM classifier for assessing osteoarthritis severity using ground reaction force measurements," *Med. Eng. Phys.*, vol. 32, no. 10, pp. 1145–1160, Dec. 2010.
- [10] D. G. Stavrakoudis, G. N. Galidaki, I. Z. Gitas, and J. B. Theocharis, "A Genetic Fuzzy-Rule-Based Classifier for Land Cover Classification From Hyperspectral Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 1, pp. 130–148, Jan. 2012.
- [11] A. Gonzalez and R. Perez, "Selection of relevant features in a fuzzy genetic learning algorithm," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 31, no. 3, pp. 417–425, Jun. 2001.
- [12] E. G. Mansoori, M. J. Zolghadri, and S. D. Katebi, "SGERD: A Steady-State Genetic Algorithm for Extracting Fuzzy Classification Rules From Data," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 4, pp. 1061–1071, Aug. 2008.
- [13] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.
- [14] W. W. Cohen, "Fast Effective Rule Induction," in *In Proceedings of the Twelfth International Conference on Machine Learning*, Lake Tahoe, California, USA, 1995, pp. 115–123.
- [15] W. W. Cohen and Y. Singer, "A Simple, Fast, and Effective Rule Learner," in *In Proceedings of the Sixteenth National Conference on Artificial Intelligence*, Orlando, Florida, USA, 1999, pp. 335–342.
- [16] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. Springer, 2000.

Depth Estimation from Single Face Image using Modified Differential Evolution

K. Punnam Chandar

Dept. of E.C.E

Kakatiya University

Warangal, INDIA

Email: k_punnam@yahoo.co.in

T. Satya Savithri

Dept. of E.C.E

Jawaharlal Nehru Technological University

Hyderabad, INDIA

Email: tirumalasatya@gmail.com

Abstract—In this paper Modified Differential Evolution (MoDE) is used to estimate the depth values of important face features from one non-frontal-view image. The objective function for optimization is formulated based on measurements from similarity transform. Further the depth estimation using MoDE is compared with Classical Differential Evolution, Particle Swarm Optimization and Non-Linear-Least Squares Optimization by computing the similarity metric, Pearson Linear Correlation Coefficient. Experimental results on 3D Bosphorus Database show that the performance of the MoDE in estimating the depths of the face features is similar to DE and outperforms in comparison with PSO and NLS1. All the simulations were carried out in Matlab.

Keywords—3D Shape Reconstruction; Structure-from-Motion; Differential Evolution; Candide Face Model;

I. INTRODUCTION

2D-face recognition & verification systems have been developed with more than two decades of research but exhibit well known deficiencies. Only under controlled conditions the 2D-face recognition systems achieve reasonable performance level [1], [2]. In physical and logical access control & security 2D wide angle Surveillance cameras are deployed for law enforcement. The use of wide angle cameras maximizes the viewing area at the cost of decrease in performance as the controlled conditions are hard to meet. In real time with the query image having arbitrary pose with very small face region when the person is far away from the camera the performance of 2D face recognition systems suffers dramatically.

One way to improve the Face recognition performance under arbitrary pose is to transforming the probe image into a canonical position [3]. This method relies on accurate landmark detection. The true transformation is nonlinear and subject dependent and therefore unknown. The alternative is to use multiple training images under different poses [4]. In real time multiple face images of the subject under consideration may not be immediately available, and even if available it could not be done without the quantization of the view sphere which would again lead to inaccuracies. On the other hand use of multiple face images will increase the storage and computation time required for further useful analysis. A potential method to improve the recognition performance in low quality input video data is to employ super resolution algorithms prior to face recognition. For these methods to give satisfactory performance, the low quality images should comply data constraints like visual quality and face similarity and will be

computationally burden. Therefore 3D face data acquisition and shape representation research is gaining momentum, as this is the only information that is invariant in face imaging and should constitute a solid base for robust face recognition and verification. The applications of 3D face models are envisaged for face tracking, face animation and other multimedia applications. A comprehensive review of 3D and multi-modal 3D+2D face recognition is presented in [5].

The 3D reconstruction techniques are grouped into two categories: active sensing where a structured light pattern is projected onto the face to facilitate reconstruction; and passive sensing where reconstruction is performed directly from the facial appearance in images or video.

Active sensing technology operates on the principle of projecting a structured illumination pattern into the scene to facilitate 3D reconstruction. This type of acquisition systems work on two principles: time-of-flight; and triangulation. For these systems (Minolta, Cyberware, Hamamatsu) to capture the 3D shape requires compliant subjects without dynamic events like expressions. On the other hand the high cost and speed limitation of these systems are the obvious short comings to acquire sufficient and useful data for further processing.

Passive sensing reconstructs 3D shape from 2D images [6] and video sequences [7] based on shape-from-X techniques. The potential advantages of passive techniques are reconstruction of 3D faces from image sequences with natural illumination, simultaneously acquisition of color appearance and video-rate shape capture. This way of 3D reconstruction mainly depends on the development of novel algorithms which can utilize the existing 2D surveillance databases. An efficient 3D reconstruction algorithm utilizing prior models to regularize the reconstruction can greatly enhance the recognition and verification accuracy from existing 2D acquisition systems.

During the past decade many 3D reconstruction algorithms from 2D images have been developed and can be classified in to three groups, shape-from-shading (SFS) [8]–[11], the 3D Morphable model [12]–[14], and structure from motion [6], [15], [16]. Introducing and discussing all these algorithms will be out of scope of this paper and we confine to structure from motion as this form the base of our work. Structure from motion (SFM) is a viable choice for recovering 3D shape when one or more multiple view image frames of subject are available. SFM can estimate the sparse 3D structure by inverting the effect of the orthographic projection process given

a set of 2D observations of feature points. Ullman [17] proved that four point correspondences over three views yield a unique solution to motion and structure. Tomasi [15] proved the rank-3 theorem, i.e., the rank of the observation matrix is 3 under an orthographic projection, and proposed a robust factorization algorithm to factor the observation matrix into a shape matrix and a camera motion matrix using the singular value decomposition (SVD) technique. Xirouhakis and Delopoulos [18] extracted the motion and shape parameters of a rigid 3D object by computing the rotation matrices via the eigenvalues and eigenvectors of appropriate defined 2×2 matrices, where the eigenvalues are the expression of four motion vectors in two successive transitions. Bregler [19] assumed a 3D object to be non-rigid, and the observed shapes are represented as a linear combination of a few basis shapes. Further, Torresani et al. [20], introduced a Gaussian prior to constraining the shape coefficients, and the optimization is solved using the expectation-maximization [EM] algorithm. Yaming Wang et al. [21] recently proposed 3D sparse representation of non-rigid structure in the trajectory space and the optimization is solved using L1-regularized least squares problem.

In this paper we propose 3D face reconstruction algorithm from single 2D image based on structure from motion technique of passive sensing. We formulate the 3D reconstruction as a optimization problem based on similarity transform and is solved using **Modified Differential Evolution (MoDE)** a variant of Differential Evolution.

Further to improve the accuracy of the 3D face model estimation, we have used Candide face model. The Candide is a parameterized face mask specifically developed for the modelbased coding of human faces [22]. During the past several decades, Candide has been a popular face model used in different face-related applications, because of its simplicity and public availability [6], [23] among of all the existing methods. The third version of the CANDIDE model, called CANDIDE-3, is composed of 113 vertices and 168 triangular surfaces, as shown in Fig. 2. Each vertex is represented by its 3-D coordinates. Because of advantages and ease of use, Candide face model is used in our proposed algorithm as the initial face structure for 3D face reconstruction. Experimentation is carried on 3D Bosphorus Database which provides 2D and corresponding 3D points and the depth estimation results of the proposed algorithm can be compared to the actual depth values.

The remainder part of the paper is organized as follows. In section II, shape reconstruction problem is formulated. In Section III a brief introduction of Differential Evolution and Modified Differential Evolution is presented. Experimental results are given in Section IV, and concluding remarks in

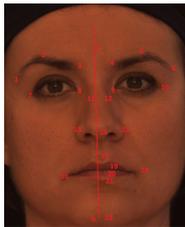


Fig. 1. Numbered Important Features in frontal view face image.



Fig. 2. Candide-3 3D face model.

Section V.

II. PROBLEM FORMULATION

In the frontal view face image mark the n 2D shape features $\sum_{i=1}^n (x_i, y_i)$ shown in Fig. 1 and in the corresponding non-frontal view face image q and is represented as coordinates $(q_{xi}, q_{yi})_{i=1,2,\dots,n}$ and is assumed marking errors free. The corresponding 3D model is represented as $(M_{xi}, M_{yi}, M_{zi})_{i=1,2,\dots,n}$ where by coordinates $(M_{xi}, M_{yi})_{i=1,2,\dots,n}$ are taken from frontal view face image and M_{zi} the depth coordinates i.e., the parameters of interest to be estimated and is initially adopted from the Candide 3D face model. The orthographic projection of the 3D model M to the 2D model is given by the similarity transform [24] and is given in Eq.1.

$$\begin{bmatrix} q_{x1} & \cdots & q_{xn} \\ q_{y1} & \cdots & q_{yn} \end{bmatrix} = k \cdot R_{2 \times 3} \cdot \begin{bmatrix} M_{x1} & M_{x2} & \cdots & M_{xn} \\ M_{y1} & M_{y2} & \cdots & M_{yn} \\ M_{z1} & M_{z2} & \cdots & M_{zn} \end{bmatrix} + \begin{bmatrix} t_{x1} & t_{x2} & \cdots & t_{xn} \\ t_{y1} & t_{y2} & \cdots & t_{yn} \end{bmatrix} \quad (1)$$

where R is the rotation matrix and is given in Eq. 2, k is the scale factor and $(t_{xi}, t_{yi})_{i=1,\dots,n}^T$ is the translations along the x and y axes. In matrix form Eq.1 can be written as in Eq.3.

$$R = \begin{bmatrix} \cos \phi & \sin \phi & 0 \\ -\sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos \psi & 0 & -\sin \psi \\ 0 & 1 & 0 \\ \sin \psi & 0 & \cos \psi \end{bmatrix} \quad (2)$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & \sin \theta \\ 0 & -\sin \theta & \cos \theta \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$$

$$q = k \cdot R_{2 \times 3} \cdot M + T \quad (3)$$

The translation term T in Eq.3 can be eliminated if the point sets to be compared are centered at the origin, i.e.,

$$\begin{aligned} q &\leftarrow q - \bar{q} \\ M &\leftarrow M - \bar{M} \end{aligned} \quad (4)$$

where \bar{q} is a $2 \times n$ matrix such that each column is $\frac{1}{n} \sum_{i=1}^n (q_{xi}, q_{yi})^T$ and \bar{M} is a $3 \times n$ matrix such that each

column is $\frac{1}{n} \sum_{i=1}^n (M_{xi}, M_{yi}, M_{zi})^T$. The Eq. 3 can be expressed as

$$q = k \cdot R_{2 \times 3} \cdot M \quad (5)$$

In the Eq. 5 it is assumed that the observation process parameters and the 3D face model fit accurately. For a given non-frontal view face image q with corresponding frontal face image given, the goal is to estimate the observation process parameters and the depths of the face model utilizing the Candide face model depths. If the estimated parameters fit the Eq.5 then the following distance will be minimum.

$$d = \|q - s \cdot R_{2 \times 3} M\|_F^2 \quad (6)$$

$$d = \left\| \begin{array}{c} q_{x1} - k(r_{11}M_{x1} + r_{12}M_{y1} + r_{13}M_{z1}) \\ q_{x2} - k(r_{11}M_{x2} + r_{12}M_{y2} + r_{13}M_{z2}) \\ \vdots \\ q_{y1} - k(r_{21}M_{x1} + r_{22}M_{y1} + r_{23}M_{z1}) \\ q_{y2} - k(r_{21}M_{x2} + r_{22}M_{y2} + r_{23}M_{z2}) \\ \vdots \end{array} \right\|_2^2 \quad (7)$$

Denoting $x = (\phi, \psi, \theta, k, M_{z1}, \dots, M_{zn})$ as the parameter vector and $f(x) = q - s \cdot R_{2 \times 3} M$, then Eq.7 can be written as

$$d = \min_x \|f(x)\|_2^2 \quad (8)$$

Z. Sun et al. [25] used Non-linear Least Squares Optimization to minimize the Eq.8 to find the solution vector \bar{x} i.e., the optimal scale, pose and depths. The NLS optimization though can produce state-of-the-art results, it requires $2n$ equations to estimate depths of n features and can also suffer from local minimum problem.

III. EVOLUTIONARY ALGORITHMS

In this paper the problem of 2D to 3D reconstruction i.e., estimation of the observation and shape parameters from Eq. 8 is performed using Modified Differential Evolution a variant of Differential Evolution.

A. Differential Evolution

Differential Evolution [28] is a recently developed population based optimization algorithm and works on the basic principle of Genetic Algorithm but differs in the no.of control parameters and the scheme of evolution of the new population. Like the GA, initial population is randomly generated with the bound constraints of the problem under consideration and fitness of the each member X_k of the initial population is evaluated. Next generation $O_k(t+1)$ is evolved from the randomly chosen three parents $X_i(t), X_j(t)$ and $X_m(t)$ from the current population based on the below operation:

$$O_k(t+1) = \begin{cases} X_m(t) + F(X_i(t) - X_j(t)), & \text{if } \text{rand}(0, 1) < Cr; \\ X_k, & \text{otherwise;} \end{cases} \quad (9)$$

where F is the scale factor and usually lies in the interval $[0, 1]$, CR is a cross over rate and is a scalar lies in the interval $[0, 1]$. The new offspring $O_k(t+1)$ is replaced with the parent if it yields a better value of the objective function in the

next generation, Otherwise, the parent under consideration is retained to the next generation. In this way the new generations are evolved to meet the desired criterion.

B. Modified Differential Evolution

In the classical DE the scale factor (F) and crossover rate (Cr) are constant for all the generations. Das et al. [26] proposed fine tuning of these parameters as the generations evolve and shown that DE with random scale factor with mean value of 0.75 and monotonic decrement of crossover rate can achieve similar result or even better result than the classical DE and coined as MoDE. The authors used the MoDE for automatic clustering of the images [27] and achieved better results in comparison to the K-Means algorithm. In this work we propose to use the MoDE with scale factor in the range $[0.1, 1]$ and retaining the monotonic decrement of the crossover rate to minimize the Eq. 8 and estimate the optimal scale, pose and depth values of the face image. It can be observed that the mean of the scale factor is 0.55. The small value of the amplification factor will constrain the depth values to a desired range while maintaining the required diversity in the population. The retaining of the monotonic decrement of the crossover rate the offsprings gets the desirable features from parents as the generations evolve.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

For the experimentation we used 3D Bosphorus Database containing the 3D and 2D coordinates for 31 subjects in different poses, expressions and occlusions. The significant advantage of this database is that it contains the 2D images with features labeled along with the corresponding 3D coordinates in raw format i.e., (x, y, z) . The availability of the 2D landmarks will avoid the possible manual marking errors. Sample images of one subject in this database are shown in Fig.3. We have used the sample images PR_D, PR_SD, PR_SU, PR_U & YR_R10 in our experimentation as these images are containing all the 22 features as shown in Fig. 1.



Fig. 3. a)Front b)PR_D c)PR_SD d)PR_SU e)PR_U f)YR_R10

As suggested by storn and price the Differential Evolution is having six strategies and are given in Table. I with two crossover schemes (Binomial and Exponential) totally DE can be used in 10 different schemes. In this work we have used all the 10 schemes of the Differential Evolution with parameter Tuning i.e., MoDE and without i.e., classical DE. To verify the accuracy of the depth values estimation we have computed the Pearson Linear Correlation Coefficient. The computed

TABLE I. SIX STRATEGIES OF DE.

Mutation Scheme	abbreviation
DE\rand\1	DE_S1
DE\local-to-best\1	DE_S2
DE\best\1 with jitter	DE_S3
DE\rand\1 with per-vector-dither	DE_S4
DE\rand\1 with per-generation-dither	DE_S5
DE\rand\1 either-or-algorithm	DE_S6

correlation coefficients for 31 subjects using MoDE & DE with strategy-2 for the five training samples with binomial crossover is shown in Fig. 4 and Fig. 5 and with exponential crossover is shown in Fig. 6 and Fig. 7 respectively. Additionally to compare the depth values estimation of the MoDE & DE we have used Particle Swarm Optimization to minimize the Eq.8 and depth values are estimated for the same five training samples and similarity metric is computed for the 31 subjects and are shown in Fig.8.

Because of the space constraint, the results of other strategies are not given, and for comparison the Pearson Correlation Coefficients of fist five subjects using strategy-1 for MoDE, DE and for PSO along with nonlinear-least-squares Optimization-1 (NLS1) are given in Table. II

TABLE II. PEARSON LINEAR CORRELATION COEFFICIENTS FOR MODE, DE, PSO & NLS1 FOR THE FIRST FIVE SUBJECTS OF THE BOSPHORUS DATABASE.

Optimization	PR_D	PR_SD	PR_SU	PR_U	YR_R10
MoDE (Bin)	0.9184	0.8823	0.8832	0.8163	0.8854
DE (Bin)	0.8970	0.8895	0.7284	0.8216	0.8824
MoDE (Exp)	0.9218	0.9040	0.8895	0.8417	0.8732
DE (Exp)	0.9091	0.9207	0.8484	0.8191	0.8680
PSO	0.7397	0.7584	0.5451	0.7858	0.7454
NLS1	0.8684	0.8023	0.2050	0.4261	0.9054

From the Table. II it can be observed that the MoDE is exhibiting the similar performance as that of the classical DE for this task but the performance is better in comparison to that of the PSO and NLS1. The mean value μ and σ of the MoDE and DE are better when compared to the PSO and NLS1. The MoDE and DE shown consistent estimation of the depth values for all the 31 subjects as is evident from Fig. 4 to Fig. 5. It can be ascertained that Evolutionary algorithm MoDE and DE are superior to PSO and NLS1 for estimating the depth values

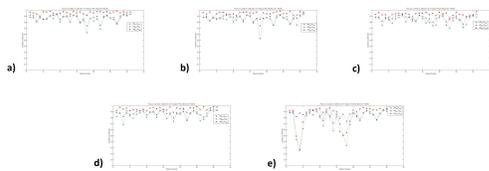


Fig. 4. Pearson Correlation Coefficients with different Training Samples a)PR_D b)PR_SD c)PR_SU d)PR_U e)YR_R10

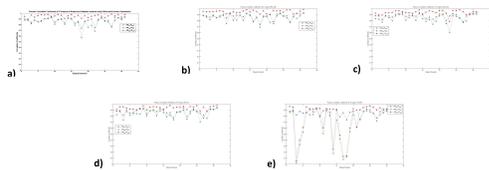


Fig. 5. Pearson Correlation Coefficients with different Training Samples a)PR_D b)PR_SD c)PR_SU d)PR_U e)YR_R10

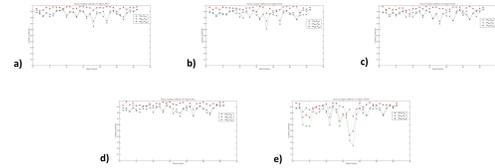


Fig. 6. Pearson Correlation Coefficients with different Training Samples a)PR_D b)PR_SD c)PR_SU d)PR_U e)YR_R10

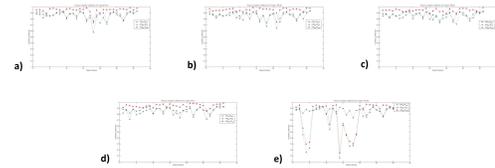


Fig. 7. Pearson Correlation Coefficients with different Training Samples a)PR_D b)PR_SD c)PR_SU d)PR_U e)YR_R10

of the face images.

V. CONCLUSION

In this work we have evaluated the performance of the DE and MoDE in estimating the depth values utilizing only one non-frontal-view face image and compared the performance with PSO and NLS1. MoDE and DE showed a consistent performance across all the 31 subjects of the Bosphorus Database for the chosen five subjects and outperformed in comparison with the PSO and NLS1. In future work different metrics will be investigated to measure the similarity of the estimated and true depths and incorporate additional prior information to make the depth estimation robust.

REFERENCES

- [1] Chellappa, Rama, Charles L. Wilson, and Saad Sirohey. "Human and machine recognition of faces: A survey." Proceedings of the IEEE 83.5 (1995): 705-741.
- [2] He, Xiaofei, et al. "Face recognition using Laplacianfaces." Pattern Analysis and Machine Intelligence, IEEE Transactions on 27.3 (2005): 328-340.
- [3] Kim, Tae-Kyun, and Josef Kittler. "Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image." Pattern Analysis and Machine Intelligence, IEEE Transactions on 27.3 (2005): 318-327.
- [4] Li, Yongmin, Shaogang Gong, and Heather Liddell. "Support vector regression and classification based multi-view face detection and recognition." Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. IEEE, 2000.
- [5] Bowyer, Kevin W., Kyong Chang, and Patrick Flynn. "A survey of approaches and challenges in 3D and multi-modal 3D+ 2D face recognition." Computer vision and image understanding 101.1 (2006): 1-15.

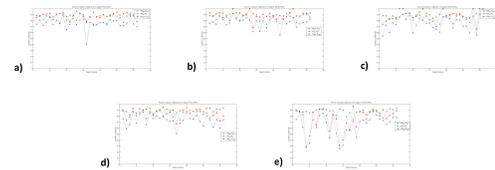


Fig. 8. Pearson Correlation Coefficients with different Training Samples a)PR_D b)PR_SD c)PR_SU d)PR_U e)YR_R10

- [6] Koo, Hei-Sheung, and Kin-Man Lam. "Recovering the 3D shape and poses of face images based on the similarity transform." *Pattern Recognition Letters* 29.6 (2008): 712-723.
- [7] Chowdhury, Amit Roy, and Rama Chellappa. "Statistical error propagation in 3d modeling from monocular video." *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on. Vol. 8. IEEE, 2003.*
- [8] Thelen, Andrea, et al. "Improvements in shape-from-focus for holographic reconstructions with regard to focus operators, neighborhood-size, and height value interpolation." *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society* 18.1 (2009): 151-157.
- [9] Casteln, Mario, and Edwin R. Hancock. "Acquiring height data from a single image of a face using local shape indicators." *Computer Vision and Image Understanding* 103.1 (2006): 64-79.
- [10] Casteln, Mario, William AP Smith, and Edwin R. Hancock. "A coupled statistical model for face shape recovery from brightness images." *Image Processing, IEEE Transactions on* 16.4 (2007): 1139-1151.
- [11] Zhang, Ruo, et al. "Shape-from-shading: a survey." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 21.8 (1999): 690-706.
- [12] Jiang, Dalong, et al. "Efficient 3D reconstruction for face recognition." *Pattern Recognition* 38.6 (2005): 787-798.
- [13] Romdhani, Sami, and Thomas Vetter. "Efficient, robust and accurate fitting of a 3D morphable model." *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE, 2003.*
- [14] Zhang, Chongzhen, and Fernand S. Cohen. "3-D face structure extraction and recognition from images using 3-D morphing and distance mapping." *Image Processing, IEEE Transactions on* 11.11 (2002): 1249-1259.
- [15] Tomasi, Carlo, and Takeo Kanade. "Shape and motion from image streams under orthography: a factorization method." *International Journal of Computer Vision* 9.2 (1992): 137-154.
- [16] Fortuna, Jeff, and Aleix M. Martinez. "Rigid structure from motion from a blind source separation perspective." *International journal of computer vision* 88.3 (2010): 404-424.
- [17] Ullman, Shimon. *The interpretation of visual motion.* Massachusetts Inst of Technology Pr, 1979.
- [18] Xirouhakis, Yiannis, and Anastasios Delopoulos. "Least squares estimation of 3D shape and motion of rigid objects from their orthographic projections." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.4 (2000): 393-399.
- [19] Bregler, Christoph, Aaron Hertzmann, and Henning Biermann. "Recovering non-rigid 3D shape from image streams." *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on. Vol. 2. IEEE, 2000.*
- [20] Torresani, Lorenzo, Aaron Hertzmann, and Christoph Bregler. "Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30.5 (2008): 878-892.
- [21] Wang, Yaming, et al. "Sparse Approximation for Nonrigid Structure from Motion." *Journal of Robotics* 501 (2015): 435385.
- [22] Ahlberg, Jrgen. "An active model for facial feature tracking." *EURASIP Journal on applied signal processing* 2002.1 (2002): 566-571.
- [23] Xie, Xudong, and Kin-Man Lam. "Elastic shape-texture matching for human face recognition." *Pattern Recognition* 41.1 (2008): 396-405.
- [24] Werman, Michael, and Daphna Weinshall. "Similarity and affine invariant distances between 2d point sets." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 17.8 (1995): 810-814.
- [25] Sun, Zhan-Li, Kin-Man Lam, and Qing-Wei Gao. "Depth estimation of face images using the nonlinear least-squares model." *Image Processing, IEEE Transactions on* 22.1 (2013): 17-30.
- [26] Das, Swagatam, Amit Konar, and Uday K. Chakraborty. "Two improved differential evolution schemes for faster global search." *Proceedings of the 7th annual conference on Genetic and evolutionary computation. ACM, 2005.*
- [27] Das, Swagatam, Ajith Abraham, and Amit Konar. "Automatic clustering using an improved differential evolution algorithm." *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 38.1 (2008): 218-237.
- [28] Storn, Rainer, Price, Kenneth. "A Simple and Efficient Heuristic for global Optimization over Continuous Spaces" *Journal of Global Optimization.*

The use of virtual laboratory works at the teaching of natural sciences subjects

Yevgeniya A. Daineko, Madina T. Ipalakova, Viktor G. Dmitriyev,

Andrey D. Giyenko, and Nazgul K. Rakhimzhanova

Abstract—The use of information computer systems into the education process is presented. The advantages and disadvantages of the virtual laboratory works at the teaching of natural sciences is given. The virtual computer laboratory for “Physics-1” subject on the base of Microsoft .NET XNA with the elements of 3D modeling was developed. It was shown, that the introduction of the modern information and computer technologies along with the traditional forms of education, allows enhancing the training the graduates.

Keywords— 3D modelling, C#, Information technologies, High Education, Microsoft .NET XNA, Physics, Virtual Laboratory work.

I. INTRODUCTION

At present time in such spheres of activities as education, science, techniques and technology the computer information systems are of great interest.

And the continuously development of techniques and technologies leading the appearance of new information systems, and in other hands to improving of existing's. In educational sphere integrating new technologies and comprehensive modernization are important tasks, on which pays great attention not only in the Republic of Kazakhstan, but also all around the world [1]. According to authors [2] the main focus of the modernization in education is the use of new information technologies, computerization of educational institutions and innovation activities of the higher educational institutions' teaching staff.

However, the integration of information technology in the educational process will be justified if they can effectively complement existing educational technology or have

The work was done under the funding of the Ministry of Education and Science of the Republic of Kazakhstan 2015 Scientific Program (No. 2622/GF4).

Ye. A. Daineko is with the International Information Technology University, Almaty, 050040 Kazakhstan (phone: +7-701-730-5195; e-mail: yevgeniya2001@gmail.com).

M. T. Ipalakova, is with the International Information Technology University, Almaty, 050040 Kazakhstan (e-mail: m.ipalakova@gmail.com).

V. G. Dmitriyev is with University of Oldenburg, Germany (e-mail: dmitriyev.viktor@gmail.com).

A. D. Giyenko is with the International Information Technology University, Almaty, 050040 Kazakhstan (e-mail: andrey.giyenko@gmail.com).

N. K. Rakhimzhanova is with the International Information Technology University, Almaty, 050040 Kazakhstan (e-mail: nazgul.rakhimzhanova@gmail.com).

additional advantages over traditional forms of learning.

For example, the use of virtual laboratory works at the teaching of natural sciences allows to complete laboratory works in physics, chemistry and biology more interesting and realistic, and also improve the quality of higher education.

II. VIRTUAL LABORATORY WORKS: ADVANTAGES AND DISADVANTAGES

Virtual laboratory works is a hardware and software system that allows students to conduct experiments without direct contact with the actual installation or in case of its absence [3].

Thus it is necessary to distinguish between such concepts as "virtual laboratory" and "virtual remote laboratory". The basis of the virtual laboratory is a computer program or complex of related programs, carrying out computer simulations of certain processes [4]. Virtual remote laboratory is a network organizational structure of several scientists' groups, which belongs to various scientific centers and interconnected in relationship of mutually beneficial cooperation, thanks to Internet [5].

Compared to traditional laboratories virtual laboratories have several advantages. Firstly, there is no need to buy expensive equipment and dangerous radioactive materials. For example, laboratory work on quantum or atomic or nuclear physics require specially equipped laboratories. Virtual laboratories also to study such phenomena as the photoelectric effect, the experience of Rutherford scattering of alpha particles, the determination of the crystal lattice by electron diffraction, Zeeman and Stark effects, nuclear reactors and others.

Secondly, there is the possibility of modeling processes that is not available in the usual laboratory. In particular, most of the classic laboratory work in molecular physics and thermodynamics are closed systems, the output of which is measured by a set of electrical quantities, from which by using the equations of electrodynamics and thermodynamics calculated required quantities. All the molecular-kinetic and thermodynamic processes in the experience, in this case remain unavailable to observation. In carrying out virtual laboratory works on these paragraphs of physics or chemistry students can to observe dynamic illustration of under studied physical and chemical phenomena and processes using animated models, that are inaccessible to observation in a real experiment. And also, at the same time of the experiment progress students can observe the construction of the

corresponding graphical dependencies of physical and chemical values.

Third, virtual laboratory works have more evident visualization of physical or chemical processes in comparison with traditional laboratory work. For example, it is possible to more detail and evidently study physical processes such as the movement of charged particles, which create an electrical current or principle of the p-n-junction. Also it is possible to penetrate in the processes taking place in a fraction of a second, or lasting for several years, for example, the study of planetary motion in the gravitational field of a central body.

Another advantage of the virtual laboratory work than traditional laboratory works is the safety issue. In particular, the use of virtual laboratory works in cases where we are working with high voltage or dangerous chemicals.

However, virtual laboratory works also have some drawbacks. The main is the absence of direct contact with the object of research, devices and equipment. It is impossible to prepare a specialist, who saw only the technical object on a computer screen. Its controversially that you will be willing to go to surgeon who formerly practiced only on a computer simulators. Therefore, the most reasonable solution is a combination of traditional and integration of virtual laboratory works in the educational process, by taking into account their strengths and weaknesses.

III. THE USE OF VIRTUAL LABORATORY WORKS AT THE TEACHING OF PHYSICS

Deep understanding of physics is possible through the study of its application to solve calculation, qualitative and experimental problems. If during lectures the students are introduced to theoretical questions, then during the laboratory lessons not only theoretical knowledge is applied, but also practical skills are formed in conducting physical measurements, processing and presenting the results.

Qualitative execution and successful defense of the results of laboratory works by students are impossible without an independent preliminary preparation for laboratory lessons. In the process of preparation for the next lesson, it is first necessary to examine the description of the performed work using this manual. However, this should not be limited only to this, as a theoretical introduction to each work cannot be regarded as sufficient minimum for a deep understanding of the physical fundamentals of the work. Therefore it is necessary for each work to read the book material according to the work subject. You cannot start working without mastering its basic theoretical positions, not having realized the logic of the measurement procedure, without being able to use measuring instruments related to this work. Getting started, the student must thoroughly understand purpose of this work, the overall work plan, i.e. sequence of operations during measurements. This is the main ground to begin the work at the interview with the teacher at the beginning of class.

To ensure high-quality and mobility of education at the International Information Technologies University an innovative project on training of students - a virtual computer laboratory for "Physics-1" subject (Fig. 1) was developed and

implemented, which includes six main laboratory works from the "Mechanics", "Molecular Physics", "Electricity and Magnetism", "Atomic and Quantum physics" sections.

During the development of the Virtual Physical Laboratory

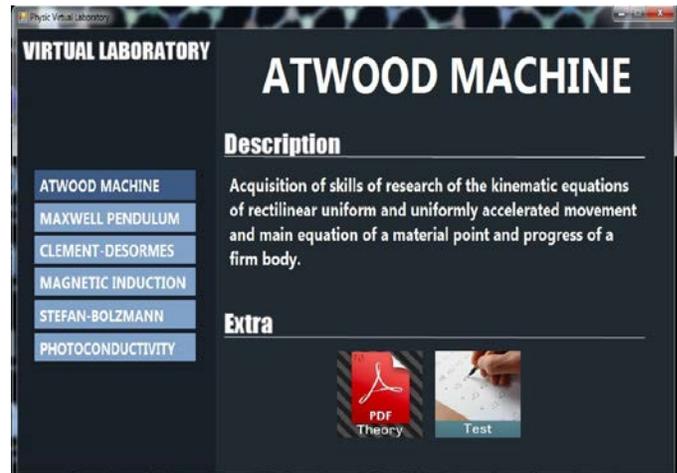


Fig. 1 fragment of the virtual laboratory on the "Physics-1" subject

3D scenes (Fig. 2), we started from the principle of economy of computer resources, ensuring the minimum system requirements, as well as realistic physical models and processes. Another important factor is the possibility of revision and editing of the program code, as well as ensuring the mutual independence of software modules for more efficient operation. In this context, it is appropriate to use an integrated development environment that supports the basic paradigms of object-oriented programming. This is an advantageous distinction of this work from existing analogues, which are difficult to various modifications and improvement.

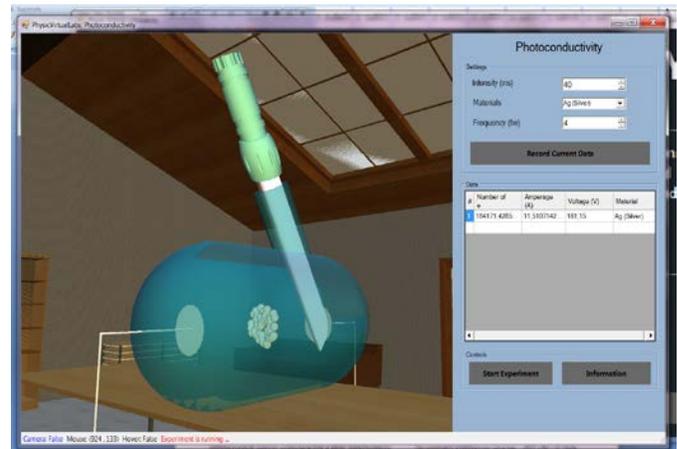


Fig. 2 3D scenes of the virtual laboratory work on the example of the investigation of the photoconductivity process

For each experiment a separate 3D model has been developed and physics engine has been implemented which computes the interaction between the objects of the model. All this has shown exact performance of the simulation of the real world objects. The models has been created using Blender and Maya 3D, and the main code has been written in C# (.NET) using the XNA 4.0 framework.

The virtual computer lab provides instructions and guidelines for the work execution, consistently built on the following form: the purpose of the work, the theoretical material, the experimental setup, the order of work execution, report. In addition, each lab contains a test that includes an assessment of the basic knowledge needed to successfully carry out the work and the final test, which aims to control the residual knowledge of the results of the laboratory work (Fig. 3).

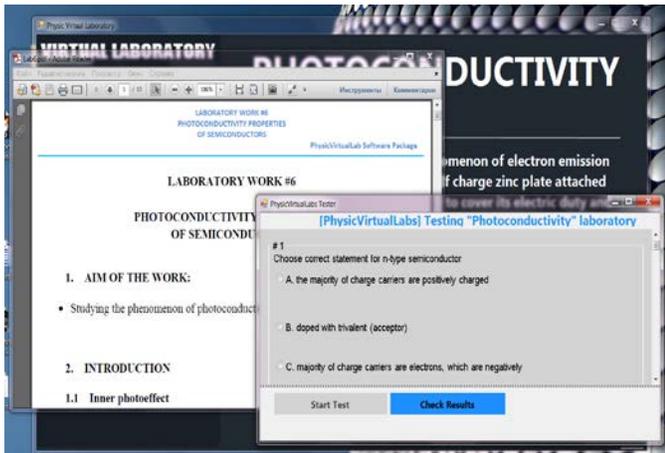


Fig. 3 instructions for the laboratory work and test example

The theoretical material in this virtual lab ushered into an electronic textbook, ie the material is in the form of hypertext that contains visual graphic and dynamic objects as well as links and tips, animated tools, reference data that help to empower students when answering the questions.

The content of the virtual lab includes a structured description of the technical facilities, which are located inside the graphic illustrations (photographs, diagrams, drawings) and hyperlinks, behind which more graphic illustrations of a similar type are "hidden", as well as video and audio clips, animations. To improve the efficiency of perception of educational material special technological methods were used, such as flash-animation, which allows you to see the experimental setup as a whole and be able to consider the smallest details. At the end of each lab are given questions for self-control and training on the material covered with brief comments, "hidden" under the hyperlink, which allows to activate the process of learning, making it interactive and help during preparation for the exams and midterms.

The students summarize the results of measurements in the form of a report. In a virtual laboratory work there are also exemplary report forms. They show what kind of tables, graphs, calculations are required in the reports. The reports must contain the conclusions drawn from the results. If necessary, the student has the right to change the form of the report for maximum clarity of the results presentation. When processing the results of measurements great attention should be paid to the calculation of measurement errors, and critical analysis of the results, to be presented in the conclusions.

Availability of reports and their presentation are the basis to pass of each laboratory work on the "Physics-1" subject.

IV. CONCLUSION

Thus, during the introduction of the advanced information technologies tools in the educational process along with the development by the future specialists of new technologies it is necessary with the help of information and communication technologies to strengthen the training of the graduate of natural science specialties, based on knowledge and understanding of the fundamental physical processes. And one of the ways is the use of the virtual laboratories.

REFERENCES

- [1] N.M. Vostrikova, "Trudy Mezhdunarodnoy konferencii," *Informatizaciya obrazovaniya*, 2008, pp. 346–348, [Online]. Available: (2008).
- [2] J. Rittinghouse, J. Ransome, *Cloud Computing: Implementation, Management, and Security*, CRC Press, 2010.
- [3] E.M. Knyazeva, "Laboratory works of new generation," *Fundamental Research*, vol. 6, 2012, pp. 587–590, [Online]. Available: (2012).
- [4] A.V. Trukhin, "Types of the virtual computer laboratories," *Open and Distance Learning*, vol. 3, 2003, pp. 12–21, [Online]. Available: (2003).
- [5] Project "Virtualnaya laboratoriya po fundamentalnym I prikladnym problemam teorii uprugosti," *International scientific-technical center* Available: <http://www.tech-db.ru/istc/db/projects.nsf/webr/1356>.

Benefits of Knowledge Engineering for e-Learning Systems

Abedl-Badeeh M. Salem and Thakaa Z. Mohamad

Abstract—Recently, knowledge engineers have begun to investigate the usage of artificial intelligence (AI) theories and approaches to develop a new generation of intelligent e-learning/training/educational systems. The main characteristics of such AI-based e-learning systems are the ability of inference, reasoning, perception, learning, and knowledge-based systems. This research demonstrates how learning systems can benefit from the innovative knowledge engineering techniques. In this paper we focus our discussion around the challenges faced by application developers and knowledge engineers designers in developing and deploying efficient and robust e-learning/education applications. Several specific examples show the applicability and efficacy of knowledge engineering techniques to e-learning in medical domain .

Keywords— knowledge engineering, e-learning, machine learning, medical e-learning, AI in education.

I. INTRODUCTION

E-learning represents a collection of e-services that employ digital media and information and communication technologies for supporting educational processes. E-learning is interdisciplinary area, encompassing many aspects of the educational technologies that cover instruction, training, teaching, learning, pedagogy, communication and collaboration. On the other side, the field of AI in education has become the most challenging area in the last several years. It includes the disciplines; cognitive and social psychology, computer science, empirical psychology, software and knowledge engineering[5] The goal of the field is to deliver computer-based systems (or knowledge-based software) which can be used in real teaching, learning and training situations.

Using AI concepts ,theories and techniques, new forms of educational software can be created that allow the computer to act as an intelligent tutor. Such AI-based intelligent tutoring system (ITS) can adjust its tutorial to the student’s knowledge, experience, strengths, and weaknesses. It may even be able to carry on a natural language dialogue. In addition, automatic generation of exercises and tests is an important feature of ITS. Moreover, intelligent e-Learning systems (IeLS) are complex to build and complex to maintain. IeLS face the knowledge-acquisition difficulty. Efficiency of IeLS is based on the

selection and determination of the appropriate knowledge representation techniques and reasoning methodologies. This paper discusses the benefits of knowledge engineering techniques for medical e-Learning systems. Also, the paper addresses the challenges facing the designing of the IeLSs.

II. KNOWLEDGE REPRESENTATION TECHNIQUES FOR INTELLIGENT E-LEARNING SYSTEMS

Knowledge can be a vague term. Research is still being done on about how can knowledge be represented so it can be manipulated and processed by a computer. From the knowledge engineering point of view, the main two components in developing an efficient and intelligent learning/educational system in any domain are the “knowledge base” and the “inference mechanism/engine”. Concerning the knowledge base, there are many knowledge representation and management techniques, e.g.; lists, trees, semantic networks, frames, scripts, production rules , cases, and ontologies[12]. The key to the success of such systems is the selection of the appropriate technique that best fits the domain knowledge and the problem to be solved. That choice is depends on the experience of the knowledge engineer. Regarding the inference engine, there are many methodologies and approaches of reasoning, e.g.; automated reasoning, case-based reasoning, commonsense reasoning, fuzzy reasoning, geometric reasoning, non-monotonic reasoning, model-based reasoning, probabilistic reasoning, causal reasoning, qualitative reasoning, spatial reasoning and temporal reasoning. In fact these methodologies receive increasing attention within the community of artificial intelligence in education . Fig 1 shows the different knowledge representation Techniques .

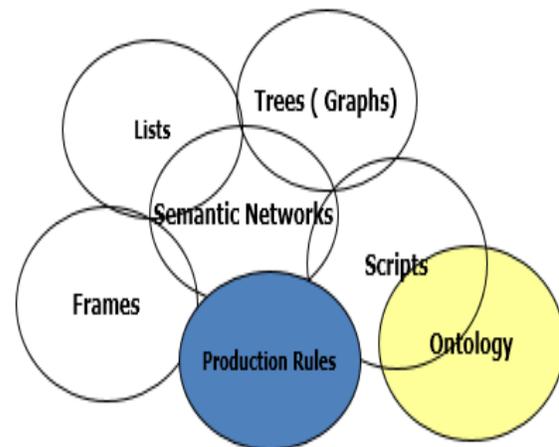


Fig 1(a) KR Techniques for static knowledge (Hierarchical Knowledge)

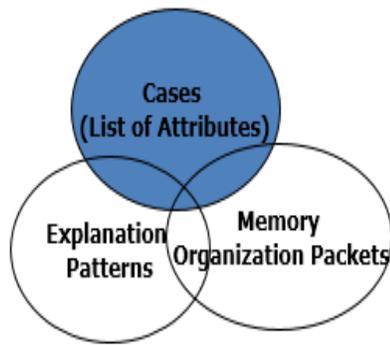


Fig 1b KR Techniques for stereotypical knowledge

A. Cases as knowledge representation for e-learning systems

The “case” is a list of features that lead to a particular outcome, e.g. the information on a patient history and the associated diagnosis. The complex form of the “case” is a connected set of sub-cases that form the problem solving task’s structure (e.g. *The design of an airplane*). Fig 2 shows the ideal components of the case. The “Case” composed of three major parts: (1) problem description, (2) solution, and (3) outcome. Problem description refers to the state of the world at the time the case is happening. Case solution refers to the stated or retrieved solution to the problem specified in the *problem description*. Case outcome defines the resulting state of the world when the solution was carried out. Depending on the case structure, the case can be used for a variety of purposes as shown in Fig 2.

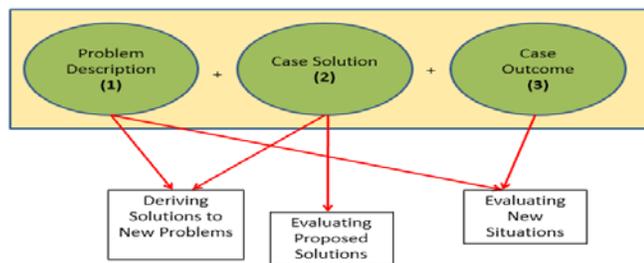


Figure 2. Depending on the case structure, the case can be used for a variety of purposes

In e-learning systems, the case can include: (a) A multi-media description of the problem, (b) A description of the correct actions to take including optimal and alternative steps, (c) A multi-media explanation of why these steps are correct, and (d) A list of methods to determine whether students correctly executed the steps. Determining the appropriate case features is the main knowledge engineering task in case-based AI software [7]. This task involves defining the terminology of the domain and gathering representative cases of problem solving by the expert. Representation of cases can be in any of several forms (*predicate, frames, scribes*). Fig 3 shows one case of an Egyptian liver cancer case description

Patient: 65-years old female not working, with nausea and vomiting.

Medical History: cancer head of pancreas .

Physical Exam: tender hepatomgaly liver, large amount of inflammatory about 3 liters, multiple liver pyogenic abscesses and large pancreatic head mass.

Laboratory Findings: total bilirubin 1.3 mg/dl, direct bilirubin 0.4 mg/dl, sgot (ast) 28 IU/L, sgpt (alt) 26 IU/L.

Figure 3: Example of an Egyptian liver cancer case description [1]

B. Ontological Approach

The term “ontology” is inherited from philosophy, in which it is a branch of metaphysics concerned with the nature of being. It began being used in AI in the 1980s, and is now frequently used by computing and information science communities. Ontological Engineering refers to the set of activities that concern the ontology development process, the ontology life cycle, the methods and methodologies for building ontologies, and the tool suites and languages that support them. During the last decade, increasing attention has been focused on ontologies [2]. At present, there are applications of ontologies with commercial, industrial, medical, academics and research focuses [13].

Ontologies' usage in learning and educational systems may be approached from various points of view: as a common vocabulary for multi-agent system, as a chain between heterogeneous educational systems, ontologies for pedagogical resources sharing or for sharing data and ontologies used to mediate the search of the learning materials on the internet [25]. The abstract specification of a system is composed of functional interconnected elements. These elements communicate using an interface and a common vocabulary. The online instructional process can be implemented successfully using artificial Intelligence techniques. Sophisticated software programs with the following features give the intelligence of the machine: adaptability, flexibility. Learning capacity, reactive capacity, autonomy, collaboration and understanding capacity. This approach enables to solve the complexity and the incertitude of the instructional systems. An intelligent learning system based on a multi-agent approach consists in a set of intelligent agents, which have to communicate. They collaborate through messages. Software agents can understand and interpret the messages due to a common ontology or the interoperability of the private ontologies. Figure 4 shows the breast cancer ontology encoded in OWL-DL format using the Protégé-OWL editing environment [3]. From this figure, it can be seen that, The breast cancers are described in terms of its symptoms, causes, stages, pathological category, diagnosis and treatment. In this context, we described causes, stages, and symptoms as

references. While diagnosis and treatment are described as medical interventions.

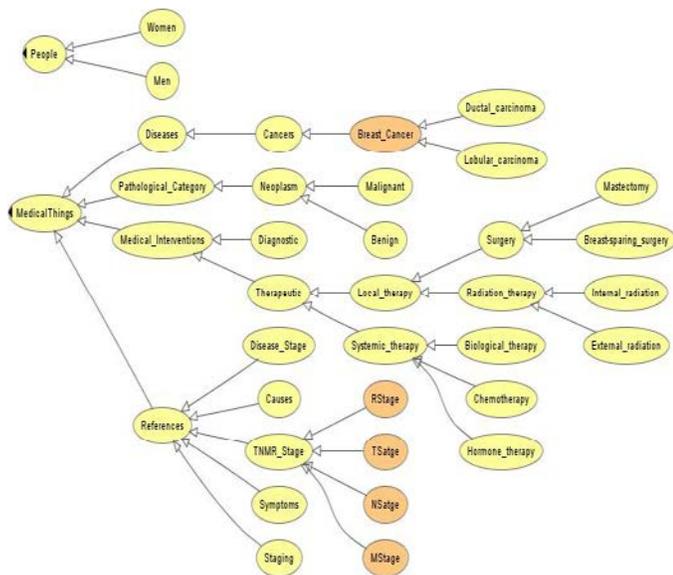


Figure 4: The Developed Breast Cancer Ontology [2].

III. DATA MINING APPROACH FOR E-LEARNING SYSTEMS

A. Data Mining Concepts

Data mining deals with the discovery of hidden knowledge, unexpected patterns and new rules from large databases [5]. Data mining (DM) and knowledge discovery (KD) is not a coherent field, it dwells upon already well established technologies including data cleaning, data preprocessing, machine learning, pattern recognition, statistics, neural networks, fuzzy sets, rough sets, clustering, etc. KD in databases process involves the following three processes; (i) using the database along with any required selection, preprocessing, sub-sampling, and transformations of it, (ii) applying DM methods (algorithms) to enumerate patterns from it, and (iii) evaluating the products of data mining to identify the subset of the enumerated patterns deemed knowledge. The data mining components of the KD process is concerned with the algorithmic means by which patterns are extracted and enumerated from data [6]. The overall KD process includes the evaluation and possible interpretation of the mined patterns to determine which patterns can be considered new knowledge. Fundamental issues in KD arise from the very nature of databases and the objects (data) they deal with. They are characterized as follows: (a) huge amounts of data, (b) dynamic nature of data, (c) incomplete or imprecise data, (d) noisy data, (e) missing attribute values and (f) redundant or insignificant data. Table 1 shows the DM tasks and the appropriate techniques for each task.

B. Benefits of Data Mining Approach in e-Learning

Recently, researchers have begun to investigate various DM methods to help instructors and administrators to improve e-learning systems [10], [14]. These methods discover new, interesting and useful knowledge based on students' usage data. Some of the main e-learning problems or subjects to which data mining techniques have been applied are dealing with the assessment of student's learning performance, provide course adaptation and learning recommendations based on the students' learning behavior, dealing with the evaluation of learning material and educational web-based courses, provide feedback to both teachers and students of e-learning courses, and detection of atypical student's learning behavior.

In what follows, we can summarize the benefits of clustering and classification in e-learning. Clustering has been used for the following e-learning tasks:

- i. Finding clusters of students with similar learning characteristics and to promote group-based collaborative learning as well as to provide incremental learner diagnosis.
- ii. Discovering patterns reflecting user behaviors and for collaboration management to characterize similar behavior groups in unstructured collaboration spaces.
- iii. Grouping students and personalized itineraries for courses based on learning objects.
- iv. Grouping students in order to give them differentiated guiding according to their skills and other characteristics.
- v. Grouping tests and questions into related groups based on the data in the score matrix.
- vi. Grouping users based on the time-framed navigation sessions.

While, classification has been used to perform the following e-learning tasks:

- i. Discovering potential student groups with similar characteristics and reactions to a specific pedagogical strategy.
- ii. Predicting students' performance and their final grade.
- iii. Detecting students' misuse or students playing around.
- iv. Predicting the students' performance as well as to assess the relevance of the attributes involved.
- v. Grouping students as hint-driven or failure-driven and finding students' common misconceptions.
- vi. Identifying learners with little motivation and finding remedial actions in order to lower drop-out rates.
- vii. Predicting course success

Data mining is a very promising approach towards the analysis of the data of student activities and behavior which accumulated by learning management systems. Data mining techniques can enhance on-line education for the educators as well as the learners. The big challenge in this respect is, while some tools using data mining techniques to help educators and

learners are being developed, the research is still in its infancy. Most of the current data mining tools are too complex for educators to use their features go well beyond the scope of what an educator might require.

Table 1: Data Mining Tasks and Techniques

Data Mining Task	Data Mining Algorithm & Technique
Clustering	K-means
Classification	Support Vector Machines Decision Trees ,Neural Network, Rule induction , Genetic Algorithms
Regression and prediction	Support Vector Machine,Decision Trees, Rule induction, NN
Association and Link Analysis (finding correlation between items in a dataset)	Association Rule Mining
Summarization	Multivariate Visualization

IV. CASE BASED REASONING (CBR) FOR E-LEARNING SYSTEMS

A. CBR Concept

CBR means reasoning from experiences or “old cases” in an effort to solve problems, critique solutions, and explain anomalous situations[7,8]. CBR is an analogical reasoning method provides both a methodology for building case-based reasoning systems, and a cognitive model of people. It is consistent with much that psychologist have observed in the natural problem solving that people do. CBR is a preferred method of reasoning in dynamically changing situations and other situations where solutions are not clear cut . From the Psychological Point of , CBR refers to reasoning in which a human problem-solver relies on previous cases that he or she has encountered.

From the computational perspective, CBR refers to a number of concepts and techniques (e.g. data structures and algorithms) that can be used to perform the following operations: (a) record and index cases, (b) search cases to identify the ones that might be useful in solving new cases when they are presented,(c) modify earlier cases to better match new cases, and (c) synthesize new cases when they are needed.

B. Benefits of CBR approach in e-Learning Systems

The idea of CBR is becoming popular in developing intelligent eLearning systems because it automates applications that are

based on precedent or that contain incomplete causal models. Research reveals that students learn best when they are presented with examples of problem-solving knowledge and are then required to apply the knowledge to real situations. The case-memory of examples and exercises capture realistic problem-solving situations and presents them to the students as virtual simulations. On the other hand, there are several benefits where students/learners should be able to perform better using CBR methodology, e.g.,

1. With more cases available, students will be able to recognize more situations and he solutions that go with these cases include failure cases, students will be able to benefit from the failures of others.

2. Retrieval cases will allow students to better recognize what is important in a new situation. Cases indexed by experts would recall and will show the student ways of looking at a problem that he might not have the expertise for without the system.

3. Student will have access to obscure cases that they otherwise would not be able to make use of. These obscure cases can help with any of the tasks previously listed

4. During a training period CBR system provides the student with a model of the way decision making ought to be done, for example, what things ought to be considered and provides them with concrete examples on which to hang their more abstract knowledge.

5. For tasks where there is much to remember, CBR systems can augment the memories of even educators. Also, both educators and students tend to focus on too few possibilities when reasoning analogically or to focus on the wrong cases.

V. AGENT-BASED APPROACHES FOR E-LEARNING

Intelligent agents (IAs) are artificial entities that have several intelligent features, such as being autonomous, responding adequately to changes in their environment , persistently pursuing goals, flexible, robust, and social by interacting with other agents. IA mimics human interaction types, such as negotiation, coordination, cooperation, and teamwork.

IAs are defined as computer systems situated in an environment and that are able to achieve their objectives by: (i) acting autonomously, i.e. by deciding themselves what to do, and (ii) being sociable, i.e. by interacting with other software agents. Agents are often seen as incarnations of various forms of AI including machine learning, reasoning and data mining. Research interests in agent systems are spanning various topics like modeling, design and development of advanced software systems that are appealing for a number of computer applications.

During the last decade, agent technologies were proposed to enhance e-learning systems across at least two dimensions :(i) agents as a modeling and design paradigm for advanced human-computer interaction and (ii) agents for smart functional decomposition of complex systems.

Firstly, agents have been described as entities that exhibit several interesting properties that are very appealing for the

modeling and design of advanced user interfaces encountered in e-learning systems: teachers, tutors and students.

Secondly, generic agent types proven to be effective for the appropriate functional decomposition of e-learning systems. Dynamic and interoperability characteristics of agents are very suitable for supporting maintainability and extensibility of e-learning systems.

VI. CONCLUSIONS

In this concluding part, we identify some of the major open problems that must be addressed to ensure the success of developing robust intelligent e-learning systems. In summary, the development of intelligent e-Learning/educational systems is a very difficult and complex process that raises a lot of technological and research challenges that have to be addressed in an interdisciplinary way. Today's the fusion of computational intelligence and machine learning techniques with the knowledge acquisition techniques solves many of the technical problems and difficulties in designing new generation of intelligent e-Learning/educational systems. Further research however is needed to convergence the knowledge engineering, artificial intelligence, machine learning, educational technology with the web science. Such convergence will create a new generation of web-based intelligent e-learning and tutoring systems. The web based of such systems can enhance the online education/ learning/training processes through the web. On the other hand, Intelligent agents technology, as a modern version of AI, where knowledge representation is enhanced with learning and social interaction. In the current environment of global wired and wireless networks, IAs may play the role of a universal carrier of distributed AI. So, the integration of software agents approaches and educational technologies is beneficial for designing efficient, robust and intelligent e-learning systems. In addition, ensuring the success of such systems to the cloud is an important challenge.

REFERENCES

- [1] Abdel-Badeeh M. Salem, "Case Based Reasoning Technology for Medical Diagnosis", Proceedings of World Academy of Science, Engineering And Technology, CESSE, Venice, Italy, Volume 25, PP 9-13, November 2007
- [2] Abdel-Badeeh M. Salem, Mohamed Roushdy, "Case-Based and Ontology Learning Approaches for Developing e-Learning Systems", WSEAS Transactions on Information Science and Applications, Issue 6, Volume 2, pp.795-804, June 2005.
- [3] Abdrabou, E. A. M. and Salem, A. B. "A Breast Cancer Classifier based on a Combination of Case-Based Reasoning and Ontology Approach." Proc. of 2nd International Multi-conference on Computer Science and Information Technology. IMCSIT 2010, Wisla, Poland, 2010.
- [4] Cios K. J., Pedrycz, W. and Swiniarski, R. W. Data Mining Methods for Knowledge Discovery. Kluwer 1998.
- [5] George F Luger, Artificial Intelligence Structure and Strategies for Complex Problem Solving, Addison Wesley, 2005
- [6] I. H. Witten and E. Frank, Data Mining – Practical Machine Learning Tools and Techniques. 2nd ed Elsevier, 2005.
- [7] Kolonder, J., Case-Based Reasoning, Morgan Kaufmann, 1993.
- [8] Lekkas, G.P., Avouris N.M. and Viras L.G., 'Cased-based reasoning in environmental monitoring applications' Applied Artificial Intelligence, 8: 359-376, 1994.
- [9] Mazza, R., & Milani, C. Exploring usage analysis in learning systems: Gaining insights from visualisations. In Workshop on usage analysis in learning systems at 12th international conference on artificial intelligence in education, New York, USA PP. 1-6, 2005.
- [10] Romero, C., & Ventura, S. Data mining in e-learning. Southampton, UK: Wit Press 2006.
- [11] Simoudis, E., 'Using case-based retrieval for customer technical support'. IEEE Expert (October), 7-12, 1992.
- [12] Sowa, J. F. - Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA., 1999
- [13] State of the Art Report 3. Ontology Interoperability, Draft version 0.3.2, http://twiki.di.uniroma1.it/pub/Estrinfo/Materiale/3._Ontology_merging.pdf, accessed on July 18, 2009
- [14] Zai'ane, O., & Luo, J. Web usage mining for a better web-based learning environment. In Proceedings of conference on advanced technology for education, Banff, Alberta, PP. 60-64 2001.

Decision Support System for predicting Football Game result

João Gomes, Filipe Portela, Manuel Filipe Santos

Abstract — there is an increase of bookmaker's number over the last decade, leading to the conclusion that the bet houses have obtained profitability in the detriment of its users. Based in this principle arises an opportunity to explore a set of artificial intelligence techniques in order to support the user betting decision. The development of this project aims to support bookmaker's users to increase their profits on bets related to football matches, suggesting to them which bet that they should carry out (home win, draw or away win). To this, it was collected several statistical information related to football games from the Premier League. It was developed a dataset and applied data mining techniques to create a model with good predictive capability. This model was then integrated in a decision support system which allows complement the machine intelligence with human perception. The model developed allowed to have profits of 20% in relation to an initial bankroll.

Keywords— Decision Support Systems, Data Mining, Football Games Prediction, Decision Support Systems for Football Betting, Knowledge Discovery in Database, Football Bets

I. INTRODUCTION

THIS paper presents the first step of a project that is being developed in the area of Decision Support Systems using Data Mining applied to football betting.

In the last few years a growing number of bookmakers has been observed in this industry, leading to the conclusion that this area presents itself as a profitable activity for the bookmakers themselves and consequently unfavorable for its users.

The project consists in analyzing football games statistics, trying to identify patterns used in these data and then leading to a suggestion for which bet should be held in a given game.

The goals of this project are divided in two groups:

- Get models with good predictive capabilities, aimed to support the users decision in a way that they can perform the best bet in a particular football game;
- Develop a prototype of an Decision Support System that incorporates the developed models.

This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

João Gomes is with Information System Department, University of Minho, Portugal.

Filipe Portela is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (Corresponding author to provide phone: +351253510319; fax: +351253510300; e-mail: cfp@dsi.uminho.pt).

Manuel Filipe Santos, is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (e-mail: mfs@dsi.uminho.pt).

The first one can be divided into the following steps:

- Collect statistical data of football games (number of goals, number of shots, etc...);
- Make a treatment of data collected;
- Create predicting data mining models;
- Make an assessment of the metrics models;
- Choose the model that satisfies the work requirements.

The second objective can be divided into:

- Create a prototype of a Decision Support System;
- Integrating the forecasting models previously created in the prototype.

After this work it was possible to find a path in order to achieve the defined goals. In this first step logical blocks were developed combined with data mining models in order to predict the better bet. In terms of model assessment, the results were not totally satisfactory being notorious the need of adding other variables and continuing the work. Although the obtained results were not the expected, it was possible to obtain models that are able to achieve good profits. By making a model evaluation, it was also possible to obtain upper than 50% correct results (more than 1/3 of the possibilities (home win, draw or away win) and profits of 20%. The achieved results gave enormous confidence to proceed with this type of research.

This paper is divided into seven topics. The first topic presents itself as an introduction to the project made. The second is called "Background" and focuses on the structuration of the theoretical context, in such way that we can ease the understanding of the project. The third topic is called "Methodologies" and in this chapter is made a description of the methodological environment concerning the development of the carried out project. The following topic is "Intelligent Decision Support System" and describes the practical work developed. The fifth topic is titled "Discussion" and it shows the results obtained through tests made evaluating the performance of the system. In the sixth topic is stated a conclusion of the theme. Finally, the last topic discusses the work that is needed to carry out in the future.

II. BACKGROUND

A. Knowledge Discovery in Data

Knowledge Discovery in Database (KDD) is described as the process of identifying and understanding incomprehensible patterns in datasets, being these patterns valuable novelties, and potentially useful [1]. This is an organized process. It is

performed automatically with an easy exploratory analysis and modeling, occurring in a given data repository [2]. KDD is an iterative process, due the fact that it may be necessary to go back one or more steps to get the results that best suit project requirements [1]. The process is composed by nine steps that can be compressed in the next five [3]:

- Selection: In this first phase a dataset selection is made to be used and sometimes it is still necessary to create a new dataset. In this phase the objectives to be achieved are also defined and it is chosen what data will be used throughout the process;
- Pre-processing: this is the phase in which is performed a cleaning and pre-processing of data;
- Transformation: At this stage the data transformation is performed in order to improve the dataset quality in order to facilitate finding data dependencies;
- Data Mining: This topic will be addressed in the topic II.B;
- Evaluation: this is the phase in which is performed an interpretation of the found patterns at the previous phase, the models are evaluated and then it is found a model that meets the objectives defined in the first phase [2].

B. Data Mining

Turban *et. al* [4] stated that the use of Data Mining (DM) from databases would become an activity that would be fundamental for organizations in the near future. This will be an important technique, so the organizations cannot waste any information regarding to their business and customers, having the risk of being overcome by its competitors.

DM is a process aiming to find patterns and interesting information in a given dataset [5].

This process was initially defined by Turban *et. al* [4] as a pattern discovery process, later they improved the concept and define DM as a process that can also be used as a way to analyze the data with the objective of increasing the efficiency and effectiveness of organizations.

The DM contains activities that can be divided into two dimensions, forecasting activities and interpretation activities [6]. In the case of this project the goal is to predict the outcome in a football game, so it will be using classification techniques.

The objective of this task was to predict the value that any random variable will assume or else estimate the probability of a future event occurs. Estimating therefore the value of a variable called 'dependent', 'target', 'response' [7].

Depending on the target class it is intended to provide the prediction of an activity. It can divided them into two specific groups: classification - if what you want is to predict the label of a class, for example "Victory", "Draw" or "Defeat", or regression if you what to want is to predict a real number "0", "1" or "2" [8].

C. Decision Support Systems

Decision Support Systems (DSS) are like an interactive computer system able to supports managers in the decision process by connecting attributes, goals and objectives, in order to solve the semi-structured and unstructured problems [9]. These are, therefore, systems aiming to support decision makers by providing alternative analysis, studying previously made

decisions and what influences these decisions had in an organizational context in order to better support the decision [10].

DSS based their development in the phases of the decision-making process. Herbert Simon [11] is the author of the methodology that gathers a greater consensus among the community. Initially he defines the decision-making process as having only three phases: Intelligence, Design and Choice. Later, together with other authors and gathering consensus they defined a fourth phase Implementation, therefore, argued that the implementation of what had been previously decided was important enough to create an individual stage for the several authors [12].

D. Similar Systems

After making a research on systems that have the objective to predict football matches results, it was possible to verify that there are platforms aiming to support the gambler decision in what will be the best bet to make in a football matches.

The DSS in development by this work has a particular feature not found in any of the existing platforms. The operation of these platforms is based on some mathematical calculations.

From the study made, the following web platforms were found:

- <http://soccervista.com/>
- <http://vitibet.com/>
- <http://pt.zulubet.com/>
- <http://www.footwin.net/>
- <http://www.predictz.com/>
- <http://www.forebet.com/>
- <https://www.statarea.com/predictions>
- <http://www.windrawwin.com/predictions/>
- <http://www.prosoccer.gr/>
- <http://spotwin.net/football-betting-system-7.html>

There is also applications for smartphones available in the AppStore or PlayStore which have the same purpose of previously described web platforms.

- KickOff – Smart Betting Made Simple
- Smart BET Prediction
- FootWin – Sports Prediction

The system that is being developed will be distinguished from these by the use of DM techniques.

III. METHODOLOGIES, MATERIAL AND METHODS

The main project follows the Design Science Research (DSR) Methodology. DSR should lead to the production of a viable artefact in the form of a building, a model, a method, or an instantiation. The main goal is to develop a technology-based solutions able to solve important and relevant business problems [13]. To complement this methodology it was used a combination of two other methodologies, the CRISP-DM and the Simon phases of decision making. For example, the first phase is composed by a combination of the Intelligence phase of decision-making and Business Understanding phase of CRISP-DM methodology. The complete methodology was presented in Table 1.

Table 1 - Combined Methodology

		Combined Methodology - DSR			
		Phase 1	Phase 2	Phase 3	Phase 4
CRISP-DM	Business Understanding	X			
	Data Understanding		X		
	Data Transformation		X		
	Modelling		X		
	Evaluation			X	
	Deployment				X
Decision-making	Intelligence	X			
	Design		X		
	Choice			X	
	Implementation				X

IV. DECISION SUPPORT SYSTEM

As previously mentioned this project was developed using a combination of three methodologies, CRISP-DM, the phases of decision-making and the DSR. At this topic is present the work developed in each phase which has been generated by the combination of the three methodologies.

A. Phase 1

In the first step was performed the identification of the problem. After making an analysis of the bookmakers market it was verified an increase of them in recent years. This situation leads to the conclusion that the profit made by them is lucrative, on the opposite only a small percentage of their users get in long-term considerable profit.

Then came the idea of creating a Decision Support System (DSS) that has the ability of help its user to make the best decision at the time it will make a bet on a certain football game, suggesting which bet must be made.

This problem corresponds to a semi-structured decision because it is necessary to complement the existing data (already collected) in the dataset with information possessed by users.

B. Phase 2

In this second phase, the project goals were defined. An analysis was made and the risks and possible restrictions that could exist for its development were determined.

The development of a DSS was made through the *Exsys Corvid* tool that would integrate data mining models developed from the *WEKA* tool, creating a system that combines intelligence machine with human perception.

Also in this phase it was carried out a dataset considered interesting and valuable for the development of the project. After making the dataset collection a detailed analysis (statistics) of it was performed in order to better understand the data that it contains and their distribution.

The data collected are detailed in Table 2. The data are from 13 seasons of the English Premier League until the season 2012/2013, which corresponds to statistical information of 4940 games.

Table 2 - Original Variables

Original Variables	Description
Date	Match Date (dd/mm/yy)
HomeTeam	Home Team
AwayTeam	Away Team
FTHG	Full Time Home Team Goals
FTAG	Full Time Away Team Goals
FTR	Full Time Result (H=Home Win, D=Draw, A=Away Win)
HTHG	Half Time Home Team Goals
HTAG	Half Time Away Team Goals
HTR	Half Time Result (H=Home Win, D=Draw, A=Away Win)
Attendance	Crowd Attendance
Referee	Name of Match Referee
HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
HHW	Home Team Hit Woodwork
AHW	Away Team Hit Woodwork
HC	Home Team Corners
AC	Away Team Corners
HF	Home Team Fouls Committed
AF	Away Team Fouls Committed
HO	Home Team Offsides
AO	Away Team Offsides
HY	Home Team Yellow Cards
AY	Away Team Yellow Cards
HR	Home Team Red Cards
AR	Away Team Red Cards
ODDS	Betting odds data from several bookmakers

After collecting the data, it was necessary to make the respective analysis. For it, an extract, transform and loading (ETL) process was designed. This process is presented in the Figure 1. In this figure is possible observe all the process since database creation, passing by the prediction phase and concluding with the dataset preparation which will allow its use in *Exsys Corvid* tool.



Figure 1 - ETL Process

This ETL is a simple process that begins with the creation of two databases, one temporary database and another one that will be later filled with all the data processed. The original dataset was uploaded to the temporary database.

Then the field "Date" was updated. This field was separated into three other fields, "Day", "Month" and "Year" in order to understand for example the weekday game.

The data presented in the dataset are mostly information that can only be obtained at the end of each game. Therefore it was necessary to identify other variables that could be known before the game starts. In this step several indicators for the home team

and for the away team were created. These indicators (table 3) are averages of the variables presented in the original dataset. Finally and before loading the final database the existing null fields were removed.

Table 3 - New Variables

New Variables	Description
AverageGoalsHomeTeam	This variable stores the goal average of the home team, in the matches disputed at home.
AverageGoalsAwayTeam	This variable stores the goal average of the away team, in the matches disputed out.
AverageShotsHomeTeam	This variable stores the average shots of the home team, in the matches disputed at home.
AverageShotsAwayTeam	This variable stores the average shots of away team, in the matches disputed out.
AverageShotsTargetHomeTeam	This variable stores the average shots on goal the home team, in the matches disputed at home.
AverageShotsTargetAwayTeam	This variable stores the average shots on goal of away team, in the matches disputed out.
HomeWinLastFive	This variable stores the number of victories in the last five games of the home team in home games.
AwayWinLastFive	This variable stores the number of victories in the last 5 games of away team in matches disputed outside.
HomeWinLastFiveConfrontation	This variable stores the number of victories in the last 5 home team's games in confrontation with the away team in home games.
AwayWinLastFiveConfrontation	This variable stores the number of victories in the last 5 games of away team matched up against the home team in matches disputed outside.
Predict	This variable stores the result predicted by the data mining model.

The next step, after loading the final database, was focused in the induction of classification models with the data obtained after them been processed and transformed. The DM models were induced using three different DM techniques: Naive Bayes (NB), Decision Trees (J48) and Support Vector Machine (LibSVM) and two sampling methods, 10-Folds Cross Validation (CV) and Percentage Split where 66% of data was

used to carry out training and 33 for testing the models

The data mining models was created and tested using a scenario. The variables used in this scenario were

- Home Team;
- Away Team;
- AverageGoalsHomeTeam;
- AverageGoalsAwayTeam;
- AverageShotsHomeTeam;
- AverageShotsAwayTeam;
- AverageShotsTargetHomeTeam;
- AverageShotsTargetAwayTeam;
- HomeWinLastFive;
- AwayWinLastFive;
- HomeWinLastFiveConfrontation;
- AwayWinLastFiveConfrontation.

The target variable was:

- FTR (Final Time Result).

Then a transformation was made to the field predicted by the models, making it a field containing only numeric data. This transformation was important because only in this way the field was recognized by *Exsys Corvid* tool. Finally the database was extracted to a text file, in order to be imported in *Exsys Corvid*.

C. Phase 3

As result is possible to get three distinct predictions (Home Win- "1", Draw- "2" and AwayWin- "3"). To analyze the models, the accuracy metric provided by Confusion Matrix was used. The six models (1 scenario x 3 techniques x 2 sampling methods) were created using all the data available. In this first phase of the project (exploration phase) it was not defined any extra scenario (combination of different variables).

The obtained confusion matrix through these models has a size of 3x3 as can be seen in the following tables (table 4 to 9). Each table is corresponding to each created model.

Table 4 - Model 1

NB	1	2	3	Accuracy
1	505	188	152	0,597633
2	140	140	169	0,311804
3	105	124	282	0,551859
Accuracy Average				0,487099

Table 5 - Model 2

J48	1	2	3	Accuracy
1	701	52	92	0,829585799
2	249	87	113	0,19376392
3	232	78	201	0,39334638
Accuracy Average				0,472232033

Table 6 - Model 3

LibSVM	1	2	3	Accuracy
1	701	52	78	0,84852071
2	251	87	11	0,249284
3	218	75	218	0,426614481
Accuracy Average				0,50813962

Table 7 - Model 4

NB	1	2	3	Accuracy
1	1477	612	395	0,594605
2	435	476	455	0,348463
3	259	422	779	0,533562
Accuracy Average				0,49221

Table 8 - Model 5

J48	1	2	3	Accuracy
1	2127	145	212	0,85628
2	797	232	337	0,169839
3	666	209	585	0,400685
Accuracy Average				0,475601

Table 9 - Model 6

LibSVM	1	2	3	Accuracy
1	2122	120	242	0,854267
2	788	229	349	0,167643
3	592	203	665	0,455479
Accuracy Average				0,492463

Table 10 presents an overview of the achieved results for each model, the sampling method and algorithm used.

Table 10 - Models Evaluation

Model	Sampling Method	Algorithm	Accuracy
Model 1	Percentage Split	NaiveBayes	0.487
Model 2	Percentage Split	J48	0.472
Model 3	Percentage Split	LibSVM	0.508
Model 4	10-Folds CV	NaiveBayes	0.492
Model 5	10-Folds CV	J48	0.476
Model 6	10-Folds CV	LibSVM	0.492

The accuracy obtained was identical in all models, being model 3 the best. So model 3 was chosen with an accuracy of 50.8%. Even though this value be not high, it is better than the 33% (random probability). This work represents an early stage of the project which leads to believe that it is a great starting point.

D. Phase 4

The goal of this phase was to ensure that what it was defined in the previous phases it is applied. In this phase was performed all the development process from the data collection, data treatment and data transformation. Then the decision support system was developed.

In this task the first prototype of the project turned up. The tool *Exsys Corvid* was used to develop the prototype.

In this phase several system development designs were tested. Below it is presented the "models" which allowed to achieve best results (user profits).

The first step was used to import the data obtained in the ETL process. The variables that the system would use were also defined. Besides the already existing variables, four new variables were defined. Two for each team, because in the moment of decision-making, there are not game data. These data are normally only known a few moments before the games start.

For that the system did the following questions to the user:

- "Which is the present classification of the home team?"
- "Which is the present classification of the away team?"
- "How many holders, from the user's opinion, are unavailable of home team?"
- "How many holders, from the user's opinion, are unavailable of away team?"

Moreover, four logic blocks were created, one was responsible for setting the value of the variables associated with the home team (Figure 2), and the other was responsible for defining the variables associated with away team. Another block was designed to do the calculation of each team score. Finally the last block makes a relationship between the scores of each team in order to suggest the best bet for the user, in this phase the data mining models are induced. It was also created one command block to the system knowing the sequence that must take (Figure 3).



Figure 2 - Logic Block Home

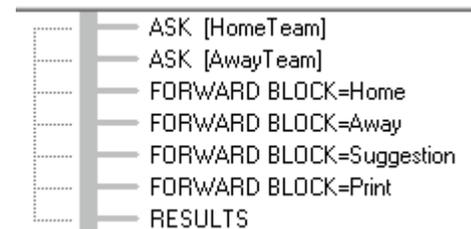


Figure 3 - Command Block

V. DISCUSSION

After the system be created, it was necessary to make an assessment of it. Regarding this fact, tests were performed at seven rounds of the English Premier League matches. They were simulated betting of € 100 in each of the 10 games by each round of matches. The obtained results are shown in Table 11.

Table 11 - System Performed Tests

Round	Percentage of Correct Bets	Return (Bets of 100€ by game)
Round 5	80 %	689 €
Round 10	30 %	-418 €
Round 15	40 %	11 €
Round 20	70 %	713 €
Round 25	60 %	480 €
Round 30	40 %	-245 €
Round 35	60 %	179 €
Total	Average = 54,29%	1409 €

These results shown that even though the model accuracy results are around 50% it was possible to have good profits. In total in this simulation was bet € 7000 and obtained a return of € 1409, about 20%, which is certainly a value to be taken into account.

In the Table 12 is presented in more detail the results obtained in the fifth round of the English Premier League in season 2013/2014.

Table 12 - Return in fifth round

	Game	Result (1,2,3)	Corvid Output (1,2,3)	Return (Bets of 100€)
Round 5	Norwich x A. Villa	3	3	220€
	Liverpool x Southampton	3	1	-100€
	Newcastle x Hull City	3	2	-100€
	West Brom x Sunderland	1	1	110€
	West Ham x Everton	3	3	140€
	Chelsea x Fulham	1	1	33€
	Arsenal x Stoke	1	1	40€
	C. Palace x Swansea	3	3	130€
	Cardiff x Tottenham	3	3	91€
	Man City x Man Utd	1	1	125€
	Total			689€

VI. CONCLUSION

This work is a first step in order to develop an Intelligent Decision Support System to predict football game results. Although the achieved results are not totally satisfactory it was possible to prove the possibility of increasing the profits on football games betting through the use of data mining classification models.

The system used data of fourteen seasons of English Premier League. Although the accuracy obtained was not very good, it was upper to 33% (probability if it was a matter of luck). The obtained profit proves that the system has enough value to continue their research and development. Scientifically this

type of models has a high level of evolution and shown to be a good option to the researchers which wants to explore this area.

In conclusion this is a project that has the potential to make the test of creating different data mining models with different types of target class to be able to make the decision support with the highest accuracy possible.

VII. FUTURE WORK

In the future the objective is to improve data mining models accuracy by increasing system reliability and consequently obtaining a higher profit. For that it will be considered the following aspects:

- To explore different types of data mining techniques;
- To create a new result variable, instead of there being three possibilities (home team win, draw or away team win), only have a chance to bet in favor of a team or against it, for example;
- To create new variables, as the weather, the players rest time;
- To use other variables which are in original dataset and which have not been used;
- Develop an Intelligent Decision Support System combining rules with the data mining engine.

REFERENCES

- [1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–53, 1996.
- [2] L. Maimon, Oded; Rokach, *Data Mining and Knowledge Discovery Handbook*, 2nd ed. 2010.
- [3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining : Towards a Unifying Framework," *Kdd*, 1996.
- [4] E. Turban, R. Sharda, and J. Aronson, "Business intelligence: a managerial approach," *Tamu-Commerce.Edu*. 2008.
- [5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.
- [6] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*. 2009.
- [7] S. Tuffery, *Data Mining and Statistics for Decision-Making*. 2011.
- [8] E. Turban, *Decision Support and Business Intelligence*, vol. 1968. 2010.
- [9] H. R. Nemat, D. M. Steiger, L. S. Iyer, and R. T. Herschel, "Knowledge warehouse: An architectural integration of knowledge management, decision support, artificial intelligence and data warehousing," *Decis. Support Syst.*, vol. 33, pp. 143–161, 2002.
- [10] J. P. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, and C. Carlsson, "Past, present, and future of decision support technology," *Decis. Support Syst.*, vol. 33, pp. 111–126, 2002.
- [11] H. A. Simon, *The New Science of Management Decision*. 1960.
- [12] H. a. Simon, *The new science of management*. 1977.
- [13] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Q.*, vol. 28, no. 1, pp. 75–105, 2004.

Prediction of potential organ donation after irreversible brain damage

Luís Torres, Filipe Portela, Manuel Filipe Santos, António Abelha, José Neves and José Machado

Abstract—Identification of patients who will die within, at least, one hour of withdrawal of life-sustaining treatment is the key to successful donation of organs after cardiac death. The accurate prediction of potential organ donation has a great importance, since the limited time window in which occur all the process demands that various tasks must be done quick and effectively. Through a set of known factors/diagnosis, it is possible to determine if a patient who suffers from irreversible brain damage may be a future candidate to organ donation with an associated degree of confidence. So in this work it was developed a prediction system, in terms of its knowledge and representation and reasoning procedures supported by a logic programming based approach to computing an artificial neural network. The factors defined and their relationships were used to identify potential organ donor.

Keywords—Prediction of potential organ donation, Degree of Confidence, Artificial Neural Network.

I. INTRODUCTION

Over the last years, advances in immunosuppressive therapeutics, better patient selection and improved technical expertise (among other factors) have decisively contributed to the success of organ transplantation, which has proven to be a successful treatment for patients with end-stage organ failure[1].

Despite its good results, this treatment has a problem with the lack of resources, i.e. organs, so that the demand far exceeds the number of available donors. The major source of organs is brain (stem) dead patients, but unfortunately (for potential organ recipients) this is not a common form of death. Furthermore, this is an undesirable outcome, since one of the goals of neurocritical care is preventing brain death from occurring [2, 3]. These results in large waiting lists, increasing everyday and considering that transplantation is often the last resort for patients with end-stage organ dysfunction, increasing the potential pool of organ donors become critical [2]. One possible way to expand the donor's group is by the use of organs from donation after cardiac death (DCD), also known as non-heart-beating donors. In general, this kind of donors are patients whose deaths occur in the context of withdrawing life-sustaining treatment (WLST).

This work was FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

Luís Torres is with Informatics Department, University of Minho, Portugal. António Abelha, José Neves and José Machado are with Algoritmi Research Centre, Braga, Portugal (email: luistorres1792@gmail.com, {Abelha, jneves, jmac}@di.uminho.pt)

Filipe Portela, University of Minho, Guimarães, Portugal. (Corresponding author to provide phone: +351253510319; fax: +351253510300; e-mail: cfp@dsi.uminho.pt).

Manuel Filipe Santos, is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (e-mail: mfs@dsi.uminho.pt).

Therapeutic failure, meaningless outcome due to poor prognosis and patient's autonomy/suffering are some of the most common reasons for WLST. The most frequent DCD candidates are patients who suffer from irreversible brain injury but do not meet criteria for brain dead diagnosis [4, 5].

Unfortunately, not all potential DCD became actual donations. The success DCD resides in the period between WLST and death. It is often associated with hypotension and poor organ perfusion that generally result in warm ischemic injury to the organs. According to British Transplantation Society and Intensive Care Society, functional warm ischaemia times vary by organ: [6]

- Liver: 30 minutes;
- Pancreas: 30 minutes;
- Lungs: 60 minutes;
- Kidney: 120 minutes;

Consequently, most DCD protocols discard organ retrieval if the patient is still alive 60 minutes after WSLT. Therefore, it has become clear the need for a model able to identify the patients who are most likely to die within 60 minutes of WLST, because only those can lead to an increase in the absolute number of available donor organs. By the other hand, it is important to clarify that the identification of a potential organ donor does not discharge a physician from treating the patient in his best interest [2, 3].

Among other benefits related to logistics and financial consequences (as reservation of operating theatre and surgical staff), a tool like this would avoid the potential organ recipient and his relatives having false expectations, situation that happens when the one hour deadline is not satisfied [4].

A few tools have been developed to predict this timing, but none is yet established as 'standard'. Between them are those from the University of Wisconsin (UW) and the United Network of Organ Sharing (UNOS). These two consist on having a numerical scale (to assign scores) and perform a trial of spontaneous respiratory rate and oxygenation when the patient is disconnected from mechanical ventilation. However, the lack information about the neurological status of the patient before WLST and the fact that they require temporary disconnection of the patient from the mechanical ventilator can be a problem. In some countries, any intervention during WLST is not directed at improving the palliative care provided is considered medically and ethically inappropriate [5]. In order to overcome this last limitation, new promising models have been developed. Coleman and her team [5] concluded that combining Glasgow Coma Scale (GCS), respiratory and haemodynamic parameters and intensivists opinion, it is possible the time from WLST to death accurately, although their results require validation in a large scale.

In other study, although not proposing any model, Suntharalingam et al [7] identified some factors that may influence the time to death. Cause of neurological injury, low blood pH, and use of inotropes prior to WLST were pointed, but younger age, higher FiO₂ and mode of ventilation were the most important variables associated with shorter time to death.

Rabinstein, Alejandro A., et al [8] build a model based on four clinical variables: absent corneal reflex, absent cough reflex, extensor or absent motor response, and higher oxygenation index. These were established as predictor variables, based on previous findings. After assigning a value to each of the variables, their sum creates a predictive score for cardiac death in patients in neurocritical state (DCD-N score). After an observational study, it was possible to translate that score in terms of probability death within 60 minutes.

By analysing this variety of models and indicators, it is clear that the key to successfully predict death after WLST lies with the identification of the correct clinic variables. Yet, the identification of the set of factors that best can characterize this problem seems something that still needs further analysis. With this article, based on some of the most important variables described, we make a start on the development of a system that can predict the potential organ donation after irreversible brain damage. We will be centred on a logic programming based approach to knowledge representation and reasoning, complemented with a computational framework based on Artificial Neural Networks.

This paper has five sections. Firstly the work and related concepts are introduced then in the second section is studied and analysed the quality of information versus the degree of confidence. In the following two section the knowledge acquired and their reasoning as also the artificial neural network are presented. Finally some conclusions are made and the future work presented.

II. QUALITY-OF-INFORMATION VERSUS DEGREE OF CONFIDENCE

Due to the growing need to offer user support in decision making processes some studies have been presented [9][10], related to the qualitative models and qualitative reasoning in Database Theory and in Artificial Intelligence (AI) research. With respect to the problem of knowledge representation and reasoning in Logic Programming (LP), a measure of the *Quality-of-Information (QoI)* of such programs has been object of some work with promising results [11], [12]. The *QoI* with respect to the extension of a predicate i will be given by a truth-value in the interval [0,1], i.e., if the information is *known (positive)* or *false (negative)* the *QoI* for the extension of *predicate_i* is 1. For situations where the information is unknown, the *QoI* is given by:

$$QoI_i = \lim_{N \rightarrow \infty} \frac{1}{N} = 0 \quad (N \gg 0) \quad (1)$$

where N denotes the cardinality of the set of terms or clauses of the extension of *predicate_i* that stand for the incompleteness under consideration. For situations where the extension of *predicate_i* is unknown but can be taken from a set of values, the *QoI* is given by:

$$QoI_i = 1/Card \quad (2)$$

where $Card$ denotes the cardinality of the *abducibles* set for i , if the *abducibles* set is disjoint. If the *abducibles* set is not disjoint, the *QoI* is given by:

$$QoI_i = \frac{1}{C_1^{Card} + \dots + C_{Card}^{Card}} \quad (3)$$

where C_{Card}^{Card} is a card-combination subset, with $Card$ elements. The next element of the model to be considered is the relative importance that a predicate assigns to each of its attributes under observation, i.e., w_i^k , which stands for the relevance of attribute k in the extension of *predicate_i*. It is also assumed that the weights of all the attribute predicates are normalized, i.e.:

$$\sum_{1 \leq k \leq n} w_i^k = 1, \forall_i \quad (4)$$

where \forall denotes the universal quantifier. It is now possible to define a predicate's scoring function $V_i(x)$ so that, for a value $x = (x_1, \dots, x_n)$, defined in terms of the attributes of *predicate_i*, one may have:

$$V_i(x) = \sum_{1 \leq k \leq n} w_i^k \times QoI_i(x)/n \quad (5)$$

It is now possible to engender all the possible scenarios of the universe of discourse, according to the information given in the logic programs that endorse the information depicted in Fig. 2, i.e., in terms of the extensions of the predicates *General Data*, *Full Outline of UnResponsive (FOUR)*, *Glasgow Coma Scores*, *DCD-N* and *Diagnosis*.

It is now feasible to rewrite the extensions of the predicates referred to above, in terms of a set of possible scenarios according to productions of the type:

$$predicate_i((x_1, \dots, x_n)) :: QoI \quad (6)$$

and evaluate the *Degree of Confidence (DoC)* given by $DoC = V_i(x_1, \dots, x_n)/n$, which denotes one's confidence in a particular term of the extension of *predicate_i*. To be more general, let us suppose that the Universe of Discourse is described by the extension of the predicates:

$$a_1(\dots), a_2(\dots), \dots, a_n(\dots) \quad (n \geq 0) \quad (7)$$

Therefore, for a given *scenario*, one may have (where \perp denotes an argument value of the type unknown; the values of the others arguments stand for themselves):

$$\left\{ \begin{array}{l} \neg a_1(x_1, y_1, z_1) \leftarrow \text{not } a_1(x_1, y_1, z_1) \\ a_1(\perp, [10,20], 15) :: 0.5 \\ \underline{[5, 10] \quad [5,30] \quad [10, 20]} \\ \text{attribute's domains for } x_1, y_1, z_1 \\ \\ \neg a_2(x_2, y_2, z_2) \leftarrow \text{not } a_2(x_2, y_2, z_2) \\ a_2([45,54], [10, 12], \perp) :: 0.65 \\ \underline{[30, 60] \quad [6, 14] \quad [2000, 6000]} \\ \text{attribute's domains for } x_2, y_2, z_2 \\ \\ \vdots \end{array} \right.$$

↓ 1st interaction: transition to continuous intervals

$$\left\{ \begin{array}{l} \neg a_1(x_1, y_1, z_1) \leftarrow \text{not } a_1(x_1, y_1, z_1) \\ a_1([5, 10], [10,20], [15, 15]) :: 0.5 \\ \underline{[5, 10] \quad [5, 30] \quad [10, 20]} \\ \text{attribute's domains for } x_1, y_1, z_1 \\ \\ \neg a_2(x_2, y_2, z_2) \leftarrow \text{not } a_2(x_2, y_2, z_2) \\ a_2([45,54], [10, 12], [2000,6000]) :: 0.65 \\ \underline{[30, 60] \quad [6, 14] \quad [2000, 6000]} \\ \text{attribute's domains for } x_2, y_2, z_2 \\ \\ \vdots \end{array} \right.$$

↓ 2nd interaction: normalization $\frac{Y - Y_{min}}{Y_{max} - Y_{min}}$

$$\left\{ \begin{array}{l} \neg a_1(x_1, y_1, z_1) \leftarrow \text{not } a_1(x_1, y_1, z_1) \\ a_1\left(\left(\frac{5-5}{10-5}, \frac{10-5}{10-5}\right), \left(\frac{10-5}{30-5}, \frac{20-5}{30-5}\right), \left(\frac{15-10}{20-10}, \frac{15-10}{20-10}\right)\right) \equiv \\ a_1([0,1], [0.2, 0.6], [0.5, 0.5]) :: 0.5 \\ \underline{[0, 1] \quad [0, 1] \quad [0, 1]} \\ \text{attribute's domains for } x_1, y_1, z_1 \\ \\ \neg a_2(x_2, y_2, z_2) \leftarrow \text{not } a_2(x_2, y_2, z_2) \\ a_2\left(\left(\frac{45-30}{60-30}, \frac{54-30}{60-30}\right), \left(\frac{10-6}{14-6}, \frac{12-6}{14-6}\right), \left(\frac{2000-2000}{6000-2000}, \frac{6000-2000}{6000-2000}\right)\right) \equiv \\ a_2([0.5, 0.8], [0.5, 0.75], [0, 1]) :: 0.65 \\ \underline{[0, 1] \quad [0, 1] \quad [0, 1]} \\ \text{attribute's domains for } x_2, y_2, z_2 \\ \\ \vdots \end{array} \right.$$

The *Degree of Confidence (DoC)* was evaluated using the equation $DoC = \sqrt{1 - \Delta l^2}$, as it is illustrated in Fig. 1. Here Δl stands for the length of the arguments' intervals, once normalized.

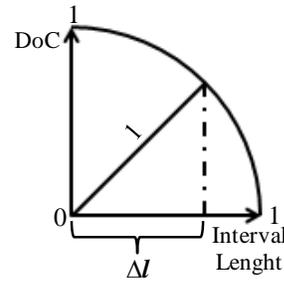


Fig. 1 Evaluation of *Degree of Confidence*

Below, one has the expected representation of the universe of discourse, where all the predicates' arguments are nominal. They speak for one's confidence that the unknown values of the arguments fit into the correspondent intervals referred to above.

$$\left\{ \begin{array}{l} \neg a_1(x_1, y_1, z_1) \leftarrow \text{not } a_1(x_1, y_1, z_1) \\ a_1(0, 0.9, 1) :: 0.5 \\ [0, 1] [0, 1] [0, 1] \\ \\ \neg a_2(x_2, y_2, z_2) \leftarrow \text{not } a_2(x_2, y_2, z_2) \\ a_2(0.9, 0.6, 0) :: 0.65 \\ [0, 1] [0, 1] [0, 1] \\ \\ \vdots \end{array} \right.$$

III. KNOWLEDGE REPRESENTATION AND REASONING

Many approaches for knowledge representation and reasoning have been proposed using the *Logic Programming (LP)* paradigm, namely in the area of Model Theory [13]–[15], and Proof Theory [16], [17]. We follow the proof theoretical approach and an extension to the *LP* language, to knowledge representation and reasoning. An *Extended Logic Program (ELP)* for short) is a finite set of clauses in the form:

$$p \leftarrow p_1, \dots, p_n, \text{not } q_1, \dots, \text{not } q_m \tag{8}$$

$$?(p_1, \dots, p_n, \text{not } q_1, \dots, \text{not } q_m) \quad (n, m \geq 0) \tag{9}$$

Where? There is a domain atom denoting falsity, the p_i , q_j , and p are classical ground literals, i.e., either positive atoms or atoms preceded by the classical negation sign \neg [17]. Under this representation formalism, every program is associated with a set of adducibles [15][18], given here in the form of exceptions to the extensions of the predicates that make the program. Once again, Logic Programming (LP) has emerged as an attractive formalism for knowledge representation and reasoning tasks, introducing an efficient search mechanism for problem solving. Therefore, and in order to exemplify the applicability of our model, we will look at the relational database model, since it provides a basic framework that fits into our expectations [19], and is understood as the genesis of the LP approach to knowledge representation and reasoning.

Consider, for instance, the scenario where a relational database is given in terms of the extensions of the relations or predicates depicted in Fig. 2, which stands for a situation where

one has to manage information about patients in a neurocritical state. Under this scenario some incomplete data is also available. For instance, in relation General Data the use of inotropes in the third patient is unknown, while in relation to Diagnosis values for pH of the first patient range in the interval [7.25, 7.35].

In relation General Data, Ventilation Mode can be: 0 – Pressure Support; Synchronised intermittent mandatory Ventilation-1; or Pressure Control/Volume Control/Pressure regulated Volume Control – 1. In what concerns to use of inotropes: 0 – not use; and 1- used.

The relation FOUR is obtained by the sum of the row, which corresponds to each of its testable components (filled according to this scale). The relation GCS works in the same way, but don't test brain stem reflexes and has a different scale. Despite this, it is still included on this model because it is the most commonly used.

The DCD-N comes from the already mentioned Rabinstein, Alejandro A., et al model, because it provides a simple way to weight some of the most important factors: for the absence of cough reflex it is given 2 points, and for the absence of each of the other 1 point.

Finally, in the relation Diagnosis, causes of neurological injury values correspond to: 3 – intracranial trauma; 2 – intracranial haemorrhage; 1 – hypoxia; 0 – other.

Now, we may consider the extensions of the relations given in Figure 2 to populate the extension of the *potential*_{donor} predicate, given in the form:

$$potential_{donor}: Age, GCS, FOUR_{score}, DCD - N_{score}, Inotropes, Cause, pH, Opinion, FiO_2 \rightarrow \{0,1\}$$

where 0 (zero) and 1 (one) denote, respectively, the truth values *false* and *true*. It is now possible to give the extension of the predicate *potential*_{donor}, in the form:

$$\{ \neg potential(Age, GCS, FOUR, DCD - N, Ino, C, pH, Op, FiO_2) \leftarrow not\ potential(Age, GCS, FOUR, DCD - N, Ino, C, pH, Op, FiO_2)$$

$$potential(20, 3, 0, 5, 1, 3, [7.25 - 7.35], 2, 0.4) :: 1. [3,75] [3,15] [0,16] [0,5] [0,1] [0,3] [7.25,7.40] [0,2] [0.15,0.5]$$

$$potential(30, 4, 4, [0,1], 0, 0, \perp, 0, 0.25) :: 1. [3,75] [3,15] [0,16] [0,5] [0,1] [0,3] [7.25,7.40] [0,2] [0.15,0.5]$$

In this program, the first clause denotes the closure of predicate *potential*_{donor}. The following clause corresponds to two terms taken from the extension of the *potential*_{donor} relation. It is now possible to have the arguments of the predicates extensions normalized to the interval [0, 1], in order to compute one's confidence that the nominal values of the arguments under considerations fit into the intervals depicted previously. One may have:

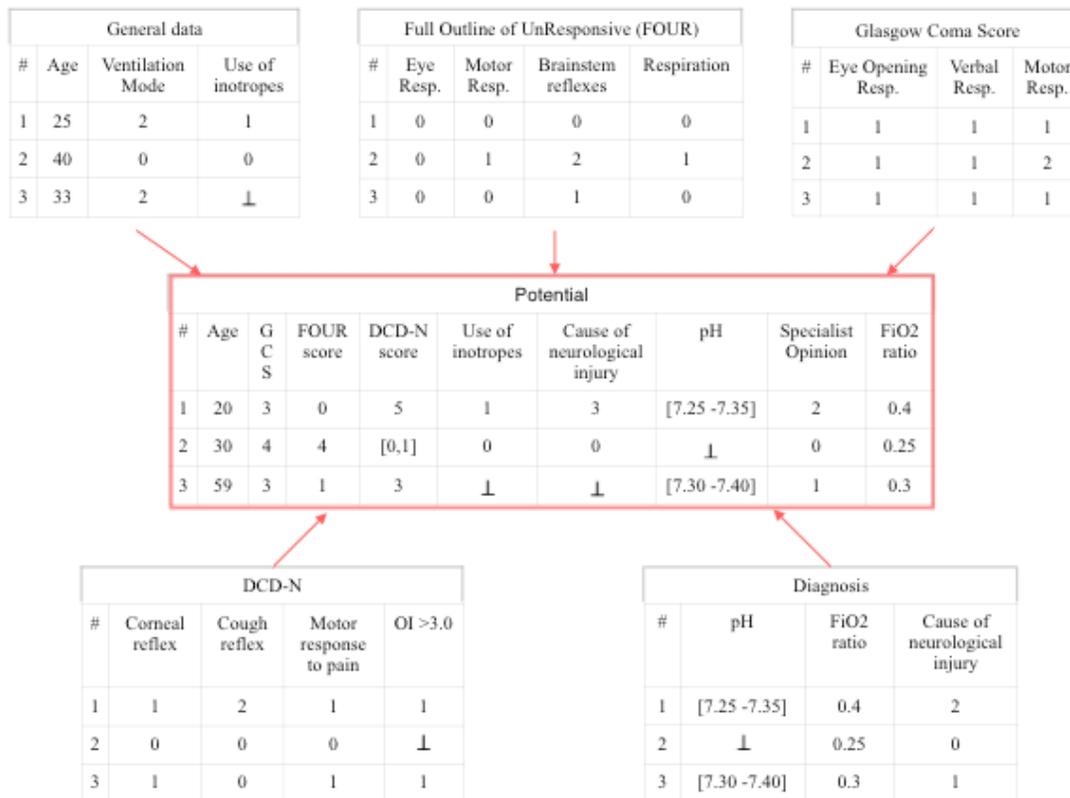


Fig. 2 An extension of the relational database model.

$$\{ potential([0.24,0.24], [0,0], [0,0], [1,1], [1,1], [1,1], [0,0.67], [1,1]$$

```
[0.71,0.71)] :: 1.
    [0,1]    [0,1] [0,1] [0,1] [0,1] [0,1] [0,1] [0,1]
    [0,1]
potential([0.375,0.375], [0.83,0.83], [0.33,0.33], [0,0.2], [0,0], [0,0],
[0,1], [0,0], [0.29,0.29]) :: 1.
    [0,1]    [0,1] [0,1] [0,1] [0,1] [0,1] [0,1] [0,1]
    [0,1]
}
```

The logic program referred to above, is now presented in the form:

```
{
¬potentialDoC(Age, GCS, FOUR, DCD - N, Ino, C, pH, Op, FiO2)
← not potentialDoC(Age, GCS, FOUR, DC - N, Ino, C, pH, Op, FiO2).
potentialDoC(1, 1, 1, 1, 1, 1, 0.74, 1, 1) :: 1.
potentialDoC(1, 1, 1, 0.98, 1, 1, 0, 1, 1) :: 1.
}
```

where its terms make the training and test sets of the following Artificial Neural Network(Figure 3).

IV. ARTIFICIAL NEURAL NETWORKS

The presented model works well to demonstrate how the information comes together to make a prediction, but it was built with the pure objective of demonstration. In order to find more reliable ways to assemble this information Artificial neural Networks (ANNs) and data mining tools can be used. Neves et al [18]–[20] demonstrated how ANNs could be successfully applied to model data and capture complex relationships between inputs and outputs. This kind of tool simulates the structure of the human brain being populated by multiple layers of neurons. As an example, let us consider the case of the third which is given in the form:

```
{
potential: (Age, GCS, FOUR, DCD - N, Ino, C, pH, Op, FiO2)
↓
potential(59, 3, 1, 3, ⊥, ⊥, [7.30 - 7.40], 1, 0.3) :: 1.
[3.75] [3.15] [0.16] [0.5] [0.1] [0.3] [7.25,7.40] [0.2] [0.15,0.5]
↓
potential([59,59], [3,3], [1,1], [ 3,3], [0,1], [0,3], [7.30
- 7.40], [1,1], [0.3, 0.3]) :: 1.
↓
potential([0.78, 0.78], [0,0], [0.06, 0.06], [0.6,0.6], [0,1], [0,1],
[0.33,0.33], [0.5, 0.5] [0.43,0.43]) :: 1.
↓
potentialDoC(1, 1, 1, 1, 0, 0, 1, 1, 1).
}
```

In Figure 3 it is shown how the normalized extremes and theirs DoC values work as inputs to the ANN. The output translates the chance of the patient death within one hour of WLST and DoC the confidence that one has on such a happening. In order to achieve good results, it is imperative to build a database of study cases that can be used to train and test the ANN.

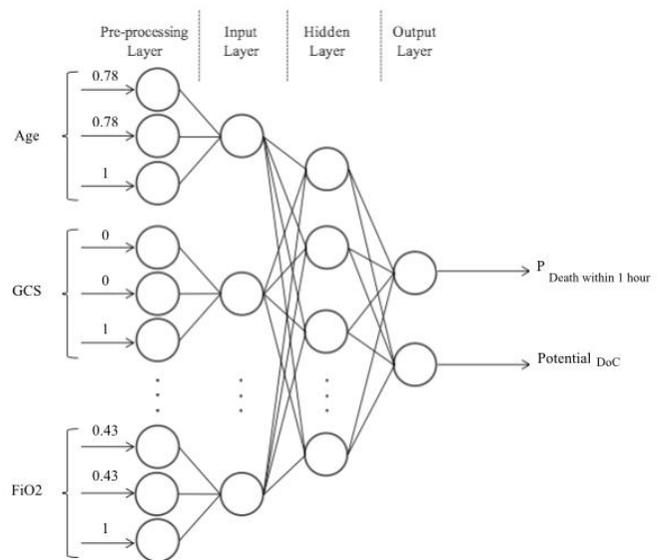


Fig. 3 ANN example for third patient

V. CONCLUSIONS AND FUTURE WORK

Identify the patients who will die in a period of 60 minutes after WLST as potential organ donors is a hard and complex task, which needs to consider many different factors with complex relations among them. All this characteristics highlight the benefits that the aid by AI techniques can bring to this field in order to achieve better prognostics.

In this work, departing from the conclusions of some good existing models, it was presented the founding of a computational framework that uses powerful knowledge representation and reasoning techniques to set the structure of the information and the associate inference mechanisms. This representation is above everything else, very versatile and capable of covering every possible instance by considering incomplete, contradictory, and even unknown data.

Future works, should study the assignment of different weights to different factors when calculating the Degree of Confidence, since the identification of the most important characteristics seems to be in the core of this problem.

ACKNOWLEDGMENTS

This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

REFERENCES

- [1] Wolfe RA, Merion RM, Roys EC, et al. Trends in organ donation and transplantation in the United States, 1998–2007. *Am J Transplant* 2009;9:869–78.
- [2] de Groot, Yorick J., et al. "Imminent brain death: point of departure for potential heart-beating organ donor recognition." *Intensive care medicine* 36.9 (2010): 1488-1494.
- [3] Kompanje, Erwin JO, Nichon E. Jansen, and Yorick J. de Groot. "In plain language": uniform criteria for organ donor recognition." *Intensive care medicine* 39.8 (2013): 1492-1494.
- [4] Wind, Jentina, et al. "Prediction of time of death after withdrawal of life-sustaining treatment in potential donors after cardiac death*." *Critical care medicine* 40.3 (2012): 766-769.
- [5] Coleman, Nicole L., E. Crowfoot, and J. L. Brieva. "Prediction of death after withdrawal of life-sustaining treatments." *Critical Care and Resuscitation* 10.4 (2008): 278.

- [6] Donation After Circulatory Dead Steering Group, "DCD Consensus Meeting Report". Reviewed:14 December 2013.
- [7] Suntharalingam, C., et al. "Time to Cardiac Death After Withdrawal of Life-Sustaining Treatment in Potential Organ Donors." *American Journal of Transplantation* 9.9 (2009): 2157-2165.
- [8] Rabinstein, Alejandro A., et al. "Prediction of potential for organ donation after cardiac death in patients in neurocritical state: a prospective observational study." *The Lancet Neurology* 11.5 (2012): 414-419.
- [9] Halpern, J.: Reasoning about uncertainty. MIT Press (2005).
- [10] Kovalerchuck, B., Resconi, G.: Agent-based uncertainty logic network. In Proc. of the WCCI 2010, IEEE, Barcelona, pp. 596–603 (2010).
- [11] Lucas, P.: Quality checking of medical guidelines through logical abduction. Proc. of AI-2003 20, pp. 309–321 (2003).
- [12] Machado, J., Abelha, A., Novais, P., Neves, J., Neves, J.: Quality of Service in healthcare units. *Int. J. Comput. Aided Eng. Technol.* 2(4), 436–449 (2010).
- [13] Kakas, A., Kowalski, R., Toni, F.: The role of abduction in logic programming. In: Gabbay, D., Hogger, C., Robinson, I. (eds) *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 5, pp. 235–324. Oxford University Press, Oxford (1998).
- [14] Gelfond, M., Lifschitz, V.: The stable model semantics for logic programming. In: Kowalski, R., Bowen, K. (eds.) *Logic programming: Proc. fifth int'l conf. and symp.*, pp. 1070–1080 (1998).
- [15] Pereira, L.M., Anh, H.T.: Evolution propection. In: Nakamatsu, K. (ed.) *Procs. first KES intl. symposium on intelligent decision technologies (KES-IDT'09)*, Himeji, Japan (2009).
- [16] Neves, J., Machado, J., Analide, C., Abelha, A., Brito, L.: The halt condition in genetic programming. In: *Lecture Notes in Artificial Intelligence*, vol. 4874. *Progress in Artificial Intelligence*. Springer, Berlin (2007).
- [17] Neves, J.: A logic interpreter to handle time and negation in logic databases, in *ACM'84 Proceedings of the 1984 annual conference of the ACM on The Fifth Generation Challenge*, pp. 50-54 (1984).
- [18] Kakas, A., Kowalski, R., Toni, F.: The role of abduction in logic programming. In: Gabbay, D., Hogger, C., Robinson, I. (eds.) *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 5, pp. 235–324. Oxford University Press, Oxford (1998).
- [19] Liu, Y., Sun, M.: Fuzzy optimization BP neural network model for pavement performance assessment. In: *2007 IEEE international conference on grey systems and intelligent services*, Nanjing, China, pp. 18–20 (2007).
- [20] Sheridan, F.: A survey of techniques for inference under uncertainty. *Artificial Intelligence Review* 5(1), 89 (1991). Z. Li, *Advanced Concrete Technology*. Hoboken, New Jersey: John Wiley & Sons, 2011, ch. 1.

Semantify Educational Resources using SKOS and Learning Object Ontologies

Georgia D. Solomou, Dimitrios A. Koutsomitropoulos, Aikaterini K. Kalou and Sotirios D. Botsios

Abstract—It is well known through experience that learning material is annotated in so many diverse ways, as the sources that maintain and curate them. Semantification of metadata descriptions can often resolve interoperability issues and strengthen the knowledge value of resources. To this end, SKOS can be a solid linking point offering a standard vocabulary for thematic descriptions. Using contemporary ontology management tools, such as WebProtégé, we show how this process can be streamlined and how it can help knowledge intensive institutions, including libraries and universities, towards aligning incoming learning material or enhancing their own.

Keywords—SKOS, Thesauri, Learning Objects, Ontologies, WebProtégé.

I. INTRODUCTION

A growing number of digital repositories systems, maintained by universities, libraries, archives and other educational institutions worldwide, are responsible for the preservation and management of educational resources. A kind of educational resource that is increasingly used by such institutions in recent years is the Learning Object (LO). In the IEEE Draft Standard for Learning Object Metadata [5], a LO is defined as "*any entity – digital or non-digital – that may be used for learning, education or training*". LOs are widely purposed and/or reused as a meaningful and effective way of creating content for e-learning [13], especially within learning- and course- management systems.

Therefore, through this work we proceed with the design and adoption of a LO metadata profile, originating from the

This work has been partially supported by the project "Information System Development for Library Functional Services" of the Democritus University of Thrace, co-financed by Greece and the European Union, in the context of Operational Programme "Digital Convergence" of the National Strategic Reference Framework (NSRF) 2007-2013.

G. D. Solomou is with the High Performance Information Systems Laboratory, (HPCLab), Computer Engineering and Informatics Dpt, University of Patras, Building B, 26500, Patras-Rio, Greece (e-mail: solomou@hpclab.ceid.upatras.gr).

D. A. Koutsomitropoulos is with the High Performance Information Systems Laboratory, (HPCLab), Computer Engineering and Informatics Dpt., University of Patras, Building B, 26500, Patras-Rio, Greece (corresponding author; phone: +30 2610 996900; fax: +30 2610 969001; e-mail: kotsomit@hpclab.ceid.upatras.gr).

A. K. Kalou is with the High Performance Information Systems Laboratory, (HPCLab), Computer Engineering and Informatics Dpt, University of Patras, Building B, 26500, Patras-Rio, Greece (e-mail: kalou@hpclab.ceid.upatras.gr).

S. D. Botsios is with Dataverse Ltd, 98 G. Papandreou Str., 54655, Kalamaria, Thessaloniki, Greece (e-mail: sdm@dataverse.gr).

widely known IEEE LOM standard. The resulting profile combines terminology with the Dublin Core metadata terms specification [1] and is intended for the efficient characterization of LOs, preserved and managed by educational institutions. Our goal is not to simply create another specialized LO metadata profile, but to contribute towards knowledge discovery across digital LOs repositories, ultimately helping institutions access, maintain and enhance learning material.

To this end, we follow a "semantification" process i.e., the transformation of the textual information captured by a metadata instance into a semantically enriched and thus machine-understandable format. Ontologies are a knowledge representation technique, offering all of the necessary constructs towards this process. They constitute the pillar of the Semantic Web, allowing knowledge reuse and sharing across applications. Ontologies have long been used for many applications in the field of education [2], so their utilization for describing educational resources can have many advantages, from facilitating the design of a LO-based course to improving the discovery of educational resources.

Going a step forward, in our LO profile's ontological representation, the subject of a LO is determined to be expressed not as a mere text keyword, but as a *concept* of a thematic thesaurus. The machine readable format of a thesaurus is achieved by the exploitation of the Simple Knowledge Organization System (SKOS) standard [10]. SKOS provides a standardized way to represent thesauri – and knowledge organization systems in general – using the Resource Description Framework (RDF) [8] and the Web Ontology Language (OWL) [11].

By combining our LO ontologies with SKOS thesauri, we can ensure a semantically enhanced characterization of LOs within the context of a digital repository, thus increasing discoverability of its resources. In addition, we set the basis for cross-repository semantic interoperability.

To manage and render accessible our LO metadata schema and ontologies, as well as any thesaurus generated explicitly to be used in combination with them, we make use of the WebProtégé ontology editor [15]. WebProtégé is a lightweight, web-based tool for ontology editing that comes with useful collaborative features and allows for the publishing of its maintained ontologies. To increase its user friendliness and aid LO ontology and thesauri maintenance and accessibility by other institutions, we have also extended WebProtégé with a couple of additional features.

To give a more thorough understanding of our work, we start by describing our LO ontology schema (Section II). We then proceed by giving the main characteristics of the SKOS model and its importance in knowledge organization, presenting also two thematic thesauri expressed in this format (Section III). In Section IV we summarize the features of the Web-Protégé editor and describe our modifications on top of it. An example of a LO ontology is given in subsequent Section V. Conclusions and future work follow in last Section VI.

II. A LEARNING OBJECT ONTOLOGY SCHEMA

Although several educational metadata schemata have been proposed over time, we are based upon the IEEE LOM standard in order to build our LO metadata profile. The reason we opted for the IEEE LOM is that this standard includes "*the minimal set of attributes needed to allow LOs to be managed, located, and evaluated*" [12] and has proven to be a widely adopted and internationally recognized open standard for the description of LOs. Our LO metadata profile adopts only a subset of the IEEE LOM element set. Our ultimate goal is the creation of a schema that would be broad enough to cover the most important educational and pedagogical aspects of an educational resource handled by a digital repository, but not exhaustively analytic, so as to become awkward in use.

The ontological binding of our LO metadata profile is expressed in the *LO Ontology Schema*. Apart from those entities representing elements originating from the IEEE LOM schema, we have also declared classes, capturing notions found in the DCMI recommendation for the Dublin Core (DC) metadata terms. This correlation helps control the values of fields for LOM properties and can increase interoperability with applications that are based on DC. In particular, LOM concepts *IntendedEndUserRole*, *InteractivityType* and *TypicalLearningTime* have been defined as refinements of the DC classes *AgentClass*, *MethodOfInstruction*, *SizeOrDuration*, respectively. For the LOM specific entities the official LOM namespace has been used (<http://ltsc.ieee.org/xsd/LOM/>, prefix lom:), whereas DC classes have been declared under the namespace <http://purl.org/dc/terms/>, prefix dcterms:.

The *lom:LearningObject* class is a top class used to capture the notion of an LO, or an educational resource in general. The various characteristics of an educational resource are represented as either classes or properties in this ontological schema. The datatype properties *lom:description*, *lom:identifier*, *lom:language*, *lom:rights*, *lom:size*, and *lom:title* are used to declare a short description, a unique identifier, the LO's content language, the copyright policies, and finally LO's physical size and title, respectively. We chose to express these elements of the LOM schema as datatype- and not as object- properties given that they simply assign values to some of the resources' basic characteristics and convey no correlations among them.

The *lom:LearningResourceType* class aims at specifying the different educational types that can be assigned to LOs and it is associated with a predefined list of terms (Exercise, Experiment, Figure, Lecture, etc.). Each such term is an instance of

the *lom:LearningResourceType* class and works as filler to the object property *lom:learningResourceType*. In a similar way, concepts met in our LO metadata profile, like the groups of end-users to which a LO applies, the intended instructional context, LO's level of difficulty, average learning time, level of completeness (draft, revised or final) and type of interaction (active, expositive, etc.) are captured using the appropriate object properties *lom:intendedEndUserRole*, *lom:context*, *lom:difficulty*, *lom:typicalLearningTime*, *lom:status*, *lom:interactivityType* respectively. These properties correlate a LO with a predefined set of values, each of which is represented as an instance of the corresponding class.

Potential relationships among LOs can be captured via the object property *lom:relation*, which is used exactly to correlate instances of the *lom:LearningObject* class. In addition, we use the *dcterms:Agent* class to include any person or organization responsible for the creation (or other modifications) to an educational resource. The object property *lom:contributor* comes to implement this type of correlation.

Finally, it is important to note that the *lom:keyword* property, used in our LO profile in order to express the thematic subject of the LO's content, is represented as an object- rather than a datatype- property. Our intention is to directly correlate the subject keywords of a LO to SKOS concepts, thus increasing the value of our LO ontology when used in the context of knowledge discovery applications. A summary of the classes and properties declared in the *LO Ontology Schema*, are shown in Fig. 1.

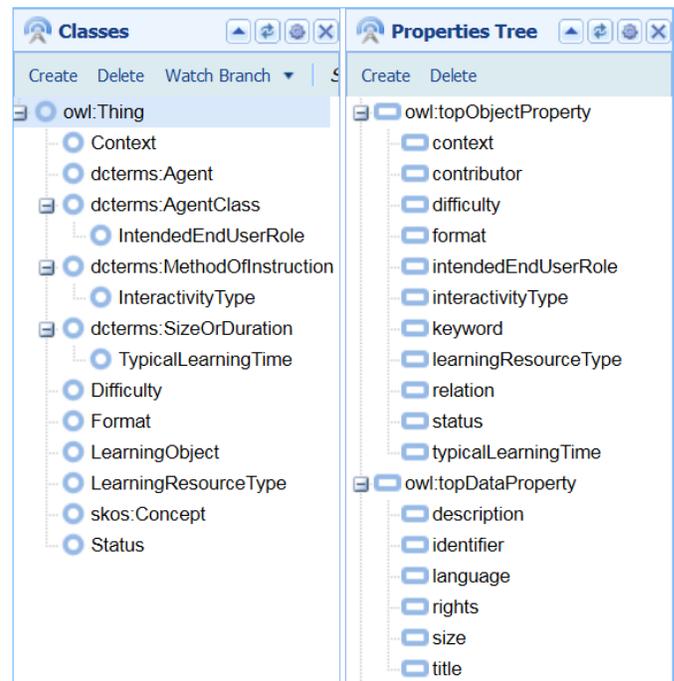


Fig. 1 Class and property hierarchies in the Learning Object Ontology Schema

Our *LO Ontology Schema* can form the basis for building more specific ontologies, targeting at the description of LOs that serve the educational purposes of various knowledge do-

mains, university courses etc. Publishing these ontologies on the Web, using a tool such as WebProtégé can significantly increase LOs management and discoverability across digital repositories. What is more, with their unique and directly accessible identifier – assigned through the *lom:identifier* datatype property – LO exposure to other discovery mechanisms, digital repositories and the Web of Linked and Open Data (LOD) [3] becomes feasible.

III. THEMATIC DESCRIPTIONS USING SKOS

SKOS is a model for expressing Knowledge Organization Systems (KOS) [4], including thesauri, in machine readable format. It provides a uniform representation of a set of terms and hence a common mechanism for the thematic indexing and retrieval of information. With the aid of SKOS, we can easily perform an integrated search against systems that are based upon controlled and structured vocabularies, such as institutional repositories and digital libraries. Additionally, as an RDF application, SKOS allows editing, publishing and interconnection of concepts on the Web, as well as their integration into other concept schemes. The terminology of SKOS has been formally expressed into RDF/OWL. An example of the SKOS structure is shown in Fig. 2.

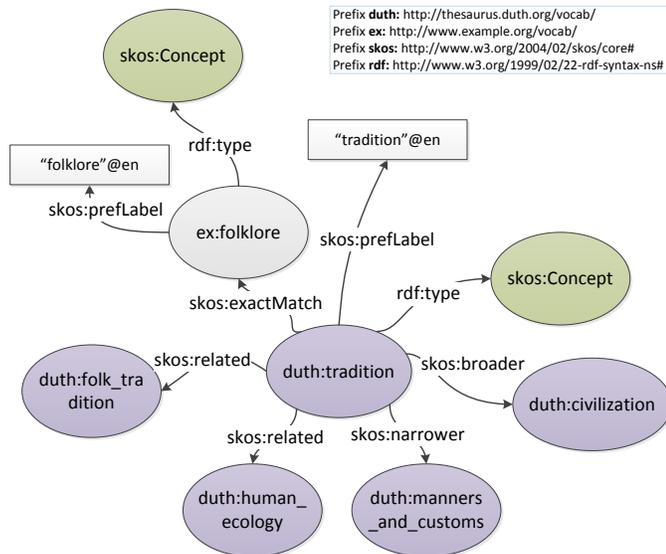


Fig. 2 Example of the structure of a SKOS concept

A. The SKOS Vocabulary

Given that SKOS is designed exactly to describe concept schemes, *concept* is its basic structural element. A SKOS concept can be viewed as a unit of knowledge, i.e., an idea or notion, an object or a class of objects and events that govern many knowledge organization systems. Therefore, *concepts* are abstract entities, which are independent of their names (i.e., the labels) used to characterize them. SKOS introduces the class *skos:Concept* to indicate that a particular term is a concept. The individuals of the *skos:Concept* class can belong to a specific concept scheme. A concept scheme is expressed through the *skos:ConceptScheme* class.

The concepts/terms of a thesaurus, when expressed in SKOS

format, are identified by URI's and assigned string labels in one or more languages. In addition, they are documented with various types of notes and interconnected with semantic relations through informal hierarchies.

Table 1 The SKOS Core Vocabulary

SKOS Term	Description
skos:Concept	An abstract idea or notion; a unit of thought
Concept Schemes	
skos:ConceptScheme	A concept scheme in which the concept is included
skos:inScheme	Relates a resource to a concept scheme in which it is included
skos:hasTopConcept	A top level concept in the concept scheme
skos:topConceptOf	Is top concept in scheme
Lexical Labels	
skos:prefLabel	The preferred lexical label for a resource, in a given language
skos:hiddenLabel	A lexical label for a resource that should be hidden when generating visual displays of the resource.
skos:altLabel	An alternative lexical label for a resource
Semantic Relations	
skos:broader	A concept that is more general in meaning
skos:narrower	A concept that is more specific in meaning
skos:broaderTransitive	Has broader transitive
skos:narrowerTransitive	Has narrower transitive
skos:related	A concept with which there is an associative semantic relationship
skos:semanticRelation	A concept related by meaning
Mapping Properties (to other concept schemes)	
skos:exactMatch	Has exact match
skos:closeMatch	Has close match
skos:broadMatch	Has broader match
skos:narrowMatch	Has narrower match
skos:relatedMatch	Has related match
skos:mappingRelation	Is in mapping relation with
Notations	
skos:notation	A string used to uniquely identify a concept within the scope of a given concept scheme
Documentation Properties	
skos:changeNote	A note about a modification to a concept
skos:definition	A statement or formal explanation of the meaning of a concept
skos:editorialNote	A note for an editor, translator or maintainer of the vocabulary
skos:example	An example of the use of a concept
skos:historyNote	A note about the past state/use/meaning of a concept
skos:note	A general note
skos:scopeNote	A note that helps to clarify the meaning and/or the use of a concept
Concept Collections	
skos:Collection	A meaningful collection of concepts
skos:OrderedCollection	An ordered collection of concepts, where both the grouping and the ordering are meaningful
skos:member	A member of a collection
skos:memberList	An RDF list containing the members of an ordered collection

To express these characteristics, the SKOS model uses a set of properties, firstly in order to define a concept itself and

secondly to relate it with other counterparts in a concept scheme. Table 1 summarizes available SKOS properties, organized into categories according to their purpose, and gives a brief description of their usage.

B. Two Thematic Terminological Thesauri

To take advantage of the potential of our LO Ontology schema, when building ontologies that capture and describe LOs, we needed a thematic thesaurus so as to directly map a LO's subject (via the *keyword* property) with SKOS concepts. These concepts would be best to originate from a standard, authoritative and controlled vocabulary rather than being arbitrary literals.

To this end, we proceeded with the creation of two thesauri – initially not in SKOS format – that cover two very common fields of knowledge: *Maths* and *Medicine*. These thesauri were actually extracted from the Thesaurus of Greek Terms, a bilingual (Greek, English) controlled vocabulary published by the National Documentation Center in Greece¹ (EKT). The latter covers a very broad field of knowledge and was created in order to facilitate libraries, museums, information centers and other institutions in Greece in characterizing and managing their digital material.

The *Maths Thesaurus* is comprised of 76 terms, making reference to 17 other related terms, whereas the *Medicine Thesaurus* contains 54 terms and makes reference to 71 additional terms. Although both of these thesauri cover specific fields of knowledge, they are generic enough and thus sufficient for the characterization of the most common subjects met in these thematic areas.

After extracting these two thesauri, our goal was to take care for their transformation into SKOS, so as to render them exploitable across different digital repositories and semantic applications. Besides, the migration of all type of knowledge organization systems into SKOS has long been recognized as a need, especially by those organizations that deal with controlled vocabularies. Some prominent examples are the Library of Congress Subject Headings (LCSH) [14] and the Food and Agriculture Organization Thesaurus² (AGROVOC).

In their initial format, both the *Maths* and the *Medicine Thesaurus* are expressed in XML syntax and follow the structure of any usual subject thesaurus, as defined by ISO 2788 [7]: they make use of hierarchical (<BT>, <NT>, <MT>), associative (<RT>) and equivalence (<UF>) relations. In addition, for each term in Greek, its English translation is provided (<ET>), as well as its correspondence to the Dewey Decimal Classification system (<dewey>).

To achieve the SKOS transformation, we implemented a mapping of the XML elements to SKOS notions, as shown in Table 2. As a result, we took the SKOS version of these two thesauri, which is in alignment with what SKOS specification defines.

¹ <http://www.ekt.gr/en/>

² <http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus>

Table 2 Mapping to SKOS elements

XML element	Function	SKOS notion
<TERM>	The described term	<skos:Concept>
<USER>	Thesaurus' owner	-
<CONTEXT>	Term's label	<skos:prefLabel lang="el">
<MT>	Microthesauri term	<skos:broaderTransitive>
<ET>	English translation	<skos:prefLabel lang="en">
<ET>	Alternative English translation	<skos:altLabel lang="en">
<BT>	Broader term	<skos:broader>
<NT>	Narrower term	<skos:narrower>
<RT>	Related term	<skos:related>
<UF>	Opposite of the Used Instead (USE) term	<skos:altLabel lang="el">
<SN>	A short description	<skos:definition>
<DEWEY>	A number indicating the correspondence to Dewey system	<skos:notation>

A snippet of a SKOS concept – belonging to the resulting SKOS version of the *Medicine Thesaurus* – can be seen in Fig. 3.

```
<skos:Concept rdf:about="http://ekt.example.org/vocab/pediatrics">
  <skos:prefLabel xml:lang="en">pediatrics</skos:prefLabel>
  <skos:prefLabel xml:lang="el">παιδιατρική</skos:prefLabel>
  <skos:inScheme rdf:resource="http://duth.example.org/vocab"/>
  <skos:broaderTransitive rdf:resource="http://ekt.example.org/vocab/
    medical_sciences"/>
  <skos:broader rdf:resource="http://ekt.example.org/vocab/medicine"/>
  <skos:related rdf:resource="http://ekt.example.org/vocab/child_psychiatry"/>
  <skos:related rdf:resource="http://ekt.example.org/vocab/children"/>
  <skos:notation rdf:datatype="http://dewey.info/schema-terms/Notation">
    618.92</skos:Notation>
</skos:Concept>
```

Fig. 3 SKOS representation of concept 'pediatrics'

IV. DEPLOYMENT ON WEBPROTEGE

WebProtégé is a free and opensource lightweight ontology editor and knowledge acquisition tool for the Web. WebProtégé allows users to create, upload, share and collaboratively edit ontologies expressed in OWL. In its current version, it is underpinned by the OWL API [6], it provides full support for OWL 2 ontologies, and comes with a simplified user interface, suitable for users with different levels of ontology expertise.

Two major features of WebProtégé that render it an appropriate tool for collaboratively deploying SKOS thesauri and ontologies and publishing them to the Web are the following:

Configurable user interface: The WebProtégé user interface is built as a portal, composed of *tabs* and *portlets* that provide independent pieces of functionality. Users can personalize UI layout, removing tabs or portlets that are not useful in their projects or adding other ones. Overall, the user interface can be configured to reflect users' OWL expertise and satisfy their projects' specific requirements.

Collaboration support: WebProtégé allows users to track changes and choose to watch entities or even whole hierarchies

of entities (branches), with the possibility to receive e-mail notifications on them. They can also have contextualized threaded discussions and notes attached to selected entities in the ontology. In addition, through an extensible access policy mechanism, users can define who may view or edit an ontology. Finally, it is possible to generate statistics of the ontology-development process.

In addition to these features, we implemented some additional facilities for WebProtégé with the intention to further enhance user's interaction with this tool and make it more convenient for editing and publishing LO ontologies and SKOS thesauri. More specifically:

- (1) An extra column, displaying the ontology's download link, has been added in the project view list of the WebProtégé home page. This link offers an explicit view of the ID that WebProtégé assigns to its projects. Additionally, it gives direct access to the corresponding WebProtégé project (ontology) and it is appropriate for use with OWL *imports* declarations.
- (2) The possibility to change the default namespace for created projects has been added. In WebProtégé this namespace is by default set to *http://webprotege.stanford.edu/*, a value that is not always desirable by project administrators. The new, implemented feature has been incorporated as an additional property option to the WebProtégé properties file and allows system administrators to customize a priori their projects' IRI prefix, based on their institutions' needs.
- (3) Similarly, another property, specifying the desired IRI suffix for each newly created entity, has been added to the same file. By setting this property, administrators can bypass system's default configuration, which is determined to use a randomly produced Universally Unique Identifier (UUID) [9] for this purpose. Now, as an alternative, they can predefine to use the entity's label (name) instead.

Although WebProtégé bears features that significantly simplify its usage, it is a tool – and not a human expert – that can't vouch for the semantic and structural correctness of the ontologies under development. Although such kind of mistakes can be eliminated using WebProtégé collaborative features, the final result is always up to the ontology expert's familiarity with OWL.

In an attempt to address this concern, we provide WebProtégé users with 'empty' templates, meant to be used as the basis for the creation of thesauri and LO ontologies. In this way an ontology expert, instead of creating a project from scratch, is encouraged to start by uploading the appropriate template. In particular, we implement a *thesaurus* template that imports the SKOS vocabulary and is used for the deployment of thematic thesauri, and a *LO Ontology* template that imports the LO ontology schema and leads to the creation of LO ontologies. The advantage of this approach is that users start building their projects having already at their disposal all necessary SKOS- or LO-specific *classes* and *properties*. As a result, they can eliminate common mistakes when building semantic corre-

lations among entities. In addition, the process of editing an ontology becomes easier, given that allowable fillers for each class are known a priori and become available through the autocomplete feature of WebProtégé. The suggested procedure workflows for deploying thesauri or LO specific projects in WebProtégé are depicted in Fig. 4.

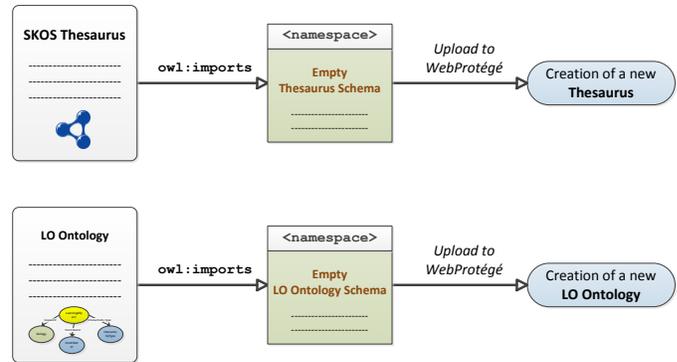


Fig. 4 Suggested procedure workflow for building a new thesaurus or a LO ontology in WebProtégé

V. AN EXAMPLE

In what follows we present an example of a LO instance, characterized using our LO metadata schema. This instance is part of a LO ontology that has been developed for semantically describing educational resources in the field of Medicine. The resulting ontology has been published through the WebProtégé editor.

The set of technical and educational characteristics of the selected LO, expressed through a set of object and datatype properties, can be seen in Fig. 5.

rdfs:label	LO_3	en	X
context	Postgraduate		X
contributor	Baharis, Konstantinos		X
contributor	Drakaki, Eleni		X
contributor	Markopoulou, Mirsini		X
difficulty	easy		X
identifier	http://hdl.handle.net/10889/3226	lang	X
intendedEndUserRole	Student		X
interactivityType	Expositive		X
keyword	Ophthalmology		X
keyword	Surgery		X
learningResourceType	Lecture		X
status	Final		X
title	Applications of Laser in Biomedical	en	X
typicalLearningTime	One_hour_to_Three_hours		X

Fig. 5 A snippet of a LO instance in WebProtégé

It is important to note that the *keyword* field of every LO has been filled using SKOS concepts coming from our Medi-

cine Thesaurus. Hence, for every LO instance captured in the ontology, the corresponding object property *keyword* has been assigned to an existing *skos:Concept* individual. This alternative for expressing a LO's subject – instead of using a mere text keyword – can lead to improved interoperability and advanced retrieval capabilities. For example, resources with content characterized by *related*, *narrower* or *broader* in meaning concepts (and captured through the corresponding SKOS properties) can also be retrieved.

Finally, through this example, it becomes evident how through its *identifier* property, the LO instance acquires a resolvable, unique identifier that provides direct access to the actual resource's location.

VI. CONCLUSIONS AND FUTURE WORK

Semantification of LO metadata can help towards having machine understandable descriptions of learning objects as well as facilitating cross-platform semantic interoperability. Starting from a LOM-based metadata profile, we have shown how to create a LO Ontology Schema and how this can be populated in order to yield semantically-enhanced descriptions of learning resources for various domains.

This Ontology Schema is further enhanced by the fact that it is possible to integrate with other ontologies, namely ones providing organization of thematic terminologies or thesauri. To foster the potential of such an approach, thesauri are expressed in SKOS format. The transformation of thesauri into SKOS is adopted by many institutions worldwide, recognizing the need to increase LOs discoverability among heterogeneous educational repositories and dissemination of knowledge.

We have demonstrated the use of WebProtégé as an environment suitable for the whole ontology lifecycle, from design to publishing, maintenance, administration and reuse. Our implemented additions on top of the system only make it more useful and convenient for this purpose.

The systematic creation and development of learning object ontologies of variable granularity (e.g., thematic-, course-oriented or other) following the LO Ontology Schema and using WebProtégé can provide educational institutions with a simple yet powerful tool for exposing their LO collections publicly. Indeed, a university or library can for example utilize the infrastructure presented in this paper in order to establish its own Learning Object Repository (LOR). In addition, it can be used as an entry point into the Web of Linked and Open Data (LOD), given the integration capabilities of the schema with SKOS or other external ontologies and datasets, while at the same time maintaining the original context and provisioning information of learning material.

As future work, semantic aware applications can be developed, that consume ontologies available through this infrastructure in various ways. For example, the thesauri we developed and maintain can seed a query expansion mechanism that searches and harvests external LORs, based on semantic matching and/or reasoning. Results from these queries can be integrated back into the LO ontologies or served to a Learning

Management System, such as e-Class, so as to widen the scope of extracurricular learning material available to students.

REFERENCES

- [1] DCMI Usage Board, "DCMI Metadata Terms," DCMI Recommendation, 2008. Available: <http://dublincore.org/documents/dcmi-terms/>
- [2] V. Devedžić, "The Setting for Semantic Web-Based Education," *Semantic Web and Education, Integrated Series in Information Systems* (12), 71-99, Springer, New York, 2006
- [3] T. Heath and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space (1st edition)," *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1), 1-136, 2011
- [4] G. Hodge, "Systems of Knowledge Organization for Digital libraries. Beyond traditional authority files," Washington, DC: the Council on Library and Information Resources, 2000. Available: <http://www.clir.org/pubs/reports/pub91/contents.html>
- [5] W. Hodgins and E. Duval, "Draft Standard for Learning Object Metadata," Institute of Electrical and Electronics Engineers, 2002. Available: http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf
- [6] M. Horridge and S. Bechhofer, "The OWL API: A Java API for Working with OWL 2 Ontologies," presented in the 6th OWL Experiences and Directions Workshop, Chantilly, Virginia, 2009
- [7] International Standards Organization, "ISO 2788:1986 Guidelines for the establishment and development of monolingual thesauri", 1986. Available: http://www.iso.org/iso/catalogue_detail.htm?csnumber=7776
- [8] G. Klyne and J. J. Carroll (eds), "Resource Description Framework (RDF): Concepts and Abstract Syntax," W3C Recommendation, 2004. Available: <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/>
- [9] P. Leach et al., "RFC 4122 - A Universally Unique Identifier (UUID) URN Namespace," Internet Engineering Task Force, Proposed Standard, 2005. Available: <http://tools.ietf.org/html/rfc4122>
- [10] A. Miles and S. Bechhofer (eds), "SKOS Simple Knowledge Organization System Reference," W3C Recommendation, 2009. Available: <http://www.w3.org/TR/skos-reference>
- [11] B. Motik, B. Parsia and P.F. Patel-Schneider (eds.), "OWL 2 Web Ontology Language XML Serialization (Second Edition)," W3C Recommendation, 2012. Available: <http://www.w3.org/TR/owl2-xml-serialization/>
- [12] S. Nair and V. Jeevan, "A Brief Overview of Metadata Formats," *DESIDOC Bulletin of Information Technology*, 24(4), pp. 3-11, 2004
- [13] P. R. Polsani, "Use and Abuse of Reusable Learning Objects", *Journal of Digital Information*, 3(4), 2003. Available: <https://journals.tdl.org/jodi/index.php/jodi/article/viewArticle/89/88>
- [14] E. Summers, A. Isaac, C. Reddin and D. Krech, "LCSH, SKOS and Linked Data", In proceedings of the International Conference on Dublin Core and Metadata Applications, 2008
- [15] T. Tudorache, C. Nyulas, N. F. Noy and M. A. Musen. "WebProtégé: A collaborative ontology editor and knowledge acquisition tool for the web," *Semantic Web Journal*, 4(1), 89-99, 2013

Big Data solutions to support Intelligent Systems and Applications

Luciana Lima, Filipe Portela, Manuel Filipe Santos, António Abelha and José Machado.

Abstract— in the last years the number of data available to be used by Intelligent Systems increased significantly. The system have now to have capabilities of storing and processing huge amount of data in real-time. With this new reality arises the Big Data concept. Big Data is much more than a big number of records stored in a database. The data can be in three formats: structured, unstructured and semi-structured. The number of Big Data solutions increase in the market, however it is difficult to understand which type of solutions are able to achieve a set of essential features: Data Integration, Data Visualization, Real-Data Analytics, Interactive Search, Text Analytics, Real-Time and Batch Processing. In order to help the researchers and professionals to have a better comprehension of the vendor's solutions and to make a better choice about what is the better solution to their Intelligent System / applications a comparative study was made. This paper present the study made and the results achieved by comparing a set of Big Data solution.

Keywords—Big Data, Intelligent Systems, Applications, Solutions, Benchmarking.

I. INTRODUCTION

THE technological evolution and consequent increased of dependence of society and organizations has led, in recent years, the exorbitant growth in the volume and variety of existing data. The McKinsey Global Institute estimates that the volume of data grow 40% per year and between 2009 and 2020 this growth will be 44 more time [1]. Every two years, the volume in all world doubles and in 2015 it will reach (approximately) 7.9 zeta bytes [1]. At the same time, market evolution requires organizations with the ability to find new ways to improve their products / services; satisfy their customers; prevent some prejudicial situations and finally, avoid the increased costs to achieve these goals.

The Big Data comes in large force, not only by the ability to process high-speed massive amounts and variety of data, but

This work was FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

Luciana Lima is with Information System Department, University of Minho, Portugal.

Filipe Portela is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (Corresponding author to provide phone: +351253510319; fax: +351253510300; e-mail: cfp@dsi.uminho.pt).

Manuel Filipe Santos, is with Algoritmi Research Centre, University of Minho, Guimarães, Portugal. (e-mail: mfs@dsi.uminho.pt).

António Abelha and José Machado are with Algoritmi Research Centre, University of Minho, Braga, Portugal. (e-mail: {Abelha,jmac}@di.uminho.pt).

also their ability to provide value to organizations who wants include Big Data in decision-making process.

The relevance of the use of Big Data by organizations of the most diverse sectors and the way how implement a solution Big Data is something that still raises questions and discussion.

Since the appearance of the Big Data buzzword some companies made its own solutions to solve their problem with complexity and volume. The success of these solutions sells Big Data as something very useful and unique. Nowadays companies are invaded by a lot of ideas, tools by all kind of suppliers.

For that reason, this paper tries to make a study of the main solution and vendors in this area. The paper display and categorize this panel of technologies with the purpose of simplify the choice. This papers presents the results achieved after a study made with the goal to find Big Data Solutions and consequently understand their features and capabilities.

Considering the amount of emerging technologies it is provided an overview of the existing technologies divided by Big Vendors (SW/HW, Cloud Deployment options) and Open Source solutions. The solutions offered by each one different and the vendors try to create connected in order to include open-source features. With this work is expected helping the decision-makers to choose the best Big Data solution without any efforts in looking for information and in understanding their main differences and values.

This paper is divided in five section. Behinds a paper introduction it is background where the main concept: Big Data and some of their features and technologies are addressed. Section three make an overview of Big Data solutions have in considerations three vectors (Commercial, Cloud and Open-Source). Then in section four it is made a summary of the results achieved during this benchmarking study. In this section the solutions were compared in two groups: technologies and commercial solutions. Finally a brief conclusion and future work are presented.

II. BACKGROUND

Big Data is a “Dataset whose size and complexity is beyond the ability of conventional tools of manage, store and analyze data”[2]. Big Data can be seen as a set of technologies capable of store, process and get value from several sources and formats of data in real-time.

Sridhar [3] considered five V's as the core of Big Data. From

the five the three with most impact are Volume, Variety, and Velocity. Veracity and Value are also important for Big Data but they are not exclusive for the new way to process data. In fact, to guarantee the Veracity and Value of data for the business is something that already is done in Business Intelligence infrastructures.

Therefore, these three V's can be defined as:

- Velocity - Data is generated, collected and processed very quickly. We fly from Batch to Real-Time and data streaming;
- Variety - Data format and sources are more diverse, much less concerned with schema or rules. Data could be collected from enterprise data warehouses, machines, web pages, etc. Structured, semi-structured or unstructured are processed as well;
- Volume - Refers to large amounts of data generated and collected every day. We jump from TB to PB measure.

As one of the most suitable technologies to storing and management big Data arises Hadoop.

Hadoop is an implementation in Java and Open Source of distributed computing used for the processing and storage of data in large scale by dividing workloads across three, five, or thousands of servers. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Companies seek deeper insights from the massive amount of structured, unstructured, semi-structured, and binary data at their disposal in order to dramatically improve business outcomes. Hadoop can help by:

- Capturing and storing all data for all business functions
- Supporting advanced analytics capabilities
- Sharing customer data quickly and generously with all those who need it
- Continuously accommodating greater data volumes and new data sources

The growth of data volume and complexity leads to the necessity of complement Hadoop with adaptable tools as is for example: Common, Avro, MapReduce, HDFS, Pig, Hive, Hbase, ZooKeeper, Sqoop.

Enterprises traditionally employ Information Technologies (IT) professionals with a package of expertise: implementation, customization, and system integration. However, Big Data deployments needs new specialist competencies – statistical and analytical – in addition to advanced IT/programming skills.

Many IT executives view open source Big Data technologies, such as Hadoop, as immature or unstable and carrying significant security and retooling risks when compared to proprietary tools

Big Data technologies enable detailed tracking and analyses of consumer profiles and behaviors, from non-traditional data sources such as social networking sites, mobile device applications, and sensors. This generates valuable business opportunities for more targeted/personalized services and cross selling. However, enterprises need to be cautious and ensure that this in-depth collection and mining of personal data does not result in privacy and compliance lapses.

III. BIG DATA SOLUTIONS

Since the appearance of the Big Data buzzword some companies made their own solutions to solve their problems with complexity and volume. The success of these solutions sells Big Data as something very useful and unique. Nowadays, companies are invaded by a lot of ideas and tools by all kind of suppliers.

For that reason, this chapter tries to display and categorize this panel of technologies with the purpose of simplifying the choice.

At this point, all the technologic offers are divided into Commercial, Cloud and Open-Source Tools. Some suppliers have their offers in commercial and open-source (sometimes free) distribution that justify their presence on both panels. Moreover, on both deployment options (Commercial, Open Source) a short summary of some technologies is provided. The Commercial section presents Big Data Solutions known as Massively Parallel Processing (MPP) tools.

Finally, in the open source section we can find the summary of Analytic Open Source Tools.

A. Commercial

The commercial or vendor-provided software is a software tool with property rights. The software is designed for sales purposes and it satisfies commercial needs. It is the model where the software developed by a commercial entity is typically licensed for a fee to a customer (either directly or through channels) in object, binary or executable code [4].

To the costumers, this kind of offer is (usually) related as a higher quality, secure and trustworthy software.

The Big Data Technology Panel (Fig. 1) exhibits commercial tools distributed in a Big Data Ecosystem that is composed by four big classes.

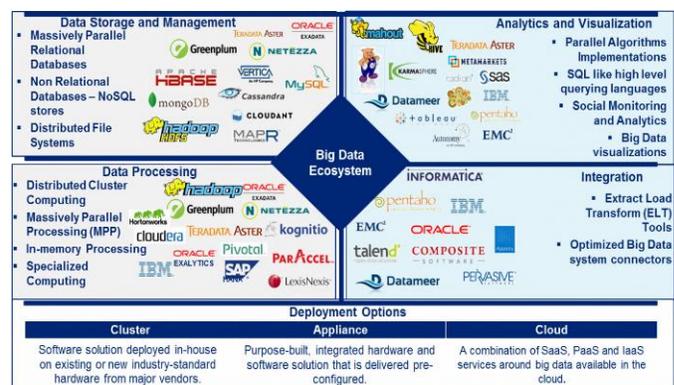


Fig. 1 - Big Data Technology Panel adapted from [5, 6]

Data Storage and Management: Technologies that have the capacity to store large volumes of different types of data. These technologies are also capable of managing and automating their process to maximize and improve the performance of their storage resources. The data storage can be in the Massively Parallel Relational Databases, Non-Relational Databases (NoSQL stores) and Distributed File Systems.

Data Processing: Technologies that are known for collecting and manipulating data to produce meaningful outputs, suitable

to be analyzed. The process could be the Distributed Cluster Computing environment, the Massively Parallel Processing (MPP) and the In-memory and Specialized Computing.

Analytics and Visualization: Technologies which are capable of providing to end-users the advanced mechanisms to explore analyze and present data in a pictorial or graphical format. In this class we found tools with Parallel Algorithms Implementations, SQL like high level querying languages, Social Monitoring and Analytics and a Big Data visualizations feature.

Integration: Technologies well suited to combine data from different sources to give an integrated view of valuable data. This process embraces extraction, cleaning, transformation tasks. The existing offer consists of Extract Load Transform (ELT) tools and Optimized Big Data system connectors.

B. Cloud

Cloud Computing it is another theme that is on top of the organizations’ minds. As a delivery model for IT services, cloud computing has the potential to enhance business agility and productivity while enabling greater efficiencies and reducing costs [7].

Although Cloud Computing is in an evolution process, it continues to mature and a growing number of enterprises are building efficient and agile cloud environments, as cloud providers continue to expand service offerings.

To manage Big Data challenges like flexibility, scalability and cost to data access, the clouds are already deployed on pools of servers, storage and networking resources and they can scale up or down according to the needs.

Cloud computing offers a cost-effective way to support big data technologies and the advanced analytics applications that can drive business value.

It will take a while for organizations to see the cloud as valid, secure and completely prepared for their needs.

As presented in the cloud vendor map (Fig. 2), there is a wide variety of cloud computing service models to provide storage, management and processing for Big Data. Organizations can leverage SaaS, PaaS or IaaS solutions depending on the needs [5].

Cloud Solution	Amazon	Microsoft	Google
Storage			
Big Data Storage	Amazon S3	HDFS	Cloud Storage (GFS)
NoSQL Store	DynamoDB	Table Storage API	AppEngine Datastore
Relational Store	MySQL or Oracle	SQL Azure	Cloud SQL
Processing			
Hosting Service	Amazon EC2	Azure Compute	AppEngine
Map Reduce Service	Elastic MapReduce	Hadoop on Azure	AppEngine Mapper API
Big Data Analytics	Elastic MapReduce	Hadoop on Azure	BigQuery



Fig. 2 - Prominent Cloud Vendors providing Big Data Solutions retrieved from [5]

C. Open Source

Open Source is a kind of offer defined as source code which must be available for everyone and all modifications made by its user can also be returned to the community.

These technologies are not necessarily free. Anyone who wants to sell an open source program, can make it however, its prices will be low whereas the development to achieve new markets will be fast.

Similar to the earlier sections of this chapter, we provide an open source technologies panel (Fig. 3) that is supplied by “native” open source companies or by renowned Big Vendors who provide some tools in an Open Source Model.



Fig. 3 - Open Source Big Data Technologies Panel retrieved from [8]

Although commercial software is usually related to higher quality and trustworthy software, Open Source Software has attracted substantial attention in the last years [4].

The same happens with cloud solutions, they represent an attractive path when we talk about resource costs but they still have not won the trust of the companies, in terms of the information systems security.

Organizations must evaluate these three options with the purpose of finding the one that best fits their commercial needs.

Be it commercial or open source tools, deployed in a cluster, cloud or appliance, the organizations must take into consideration the costs of acquisition, implementation, maintenance, upgrading and, also, the flexibility of the architecture to integrate other tools with different natures.

Finally, the security that truly depends just as much on how well the software is deployed, configured, updated and maintained, including product vulnerabilities discovered and solved through appropriate and timely updates

IV. COMPARATIVE SUMMARY

Once the investment in a Big Data project is approved and finally gathered the information about the market offers, organizations must choose the best solutions that better fits their needs. The large number of options makes it hard for organizations to decide what is best and why. This happens because each Big Data supplier sells its solution as the best one,

the most feasible, robust and scalable, amongst other features, and the organizations cannot base their decisions solely on the kind type of characteristics.

Nowadays, sharing information about it, it is the best “counseling”. Tools trial reports, vendor surveys and white papers, benchmark reports from specialists and the opinion of non-professionals (from the social media), these are the best sources to sustain our decisions. More available information simplifies the choice.

A. Analytic Open Source Tools

The table Analytic Open Source Tools (Table 1) describes some of the most used technologies presented in the panel (fig. 3). For now the table focuses only on the analytic tools because they are best known and easily accepted by companies.

This study was made having in consideration the type of features of each one has.

Table 1 - A. Analytic Open Source Tools - Summary

Key Tools	Summary	Features
Knime	Data analytics platform that allows you to perform sophisticated statistics and data mining on your data to analyze trends and predict potential results. Its visual workbench combines data access, data transformation, initial investigation, powerful predictive analytics and visualization The open integration platform offers over 1.000 modules. KNIME ¹ is also the open source data analytics platform.	Data Integration Data Mining
WEKA	WEKA stands for “Waikato Environment for Knowledge Analysis” and it is a collection of machine-learning algorithms in order to solve data mining problems. It is written in Java and thus runs on almost any modern computing platform. It supports different data mining tasks such as clustering, data pre-processing, regression, classification, feature selection as well as visualization [8].	Data Mining
Rapid Miner	RapidMiner offers data integration and analysis, analytical ETL and reporting combined in a community edition or enterprise edition. It comes with a graphical user interface for designing analysis processes. The solution offers a metadata transformation, which allows inspecting for errors during design time [8].	Data Integration Data Mining
Talend	Open source software developed by Talend has developed several big data software solutions, including Talend Open Studio for Big Data, which is a data integration tool supporting Hadoop, HDFS, Hive, Hbase and Pig. The objective is to improve the efficiency of data integration job design through a graphical development environment. Next to open source tools, Talend also sells other commercial products [8].	Data Integration
Jaspersoft	<ul style="list-style-type: none"> Jaspersoft has developed several open source tools, among others a Reporting and Analytics server, which is a standalone and embeddable reporting server. The Open Source Java Reporting Library is a reporting engine that can analyze any kind of data and produce reports in any format. Jaspersoft ETL offers a data integration engine, powered by Talend. They claim it is the world’s most used business intelligence software. 	Data Integration

¹ <http://www.knime.org/knime>

B. Big Vendor Solutions - Summary

Therefore, Table 2 presents a short summary and an objective comparison of the most popular vendors, in terms of their offers, hardware appliance and Connectors.

Table 2 - Big Vendor Solutions - Summary

Key Vendors	Offers	Hardware Appliance	Connectors
EMC Greenplum	MPP Database Hadoop distribution Chorus – search, explore, visualize, analyze Command Center	Optional- Greenplum Data Computing Appliance (DCA)- single rack expandable in quarter rack increments up to 12 racks) Optional – gNet connector	Connects to most traditional EDWs and BI / analytics applications (SAS, MicroStrategy, Pentaho)
IBM Netezza	IBM Netezza DW Appliance IBM Netezza Analytics	S-Blade servers contain multi-core Intel CPUs and IBM Netezza’s unique multi-engineFPGAs. Configuration in single and multiple racks	Data Integration with most IBM and 3rd party solutions. Working with Cloudera to bring in Hadoop connectivity
Oracle Exadata	Oracle Exadata Database Machine InfiniBand High Speed Connectivity Oracle Data Integrator	Exadata Storage Server X2-2 Exadata Database Machine X2-2 Exadata Storage Expansion Racks Exadata X2-2 Memory Expansion	Data Integration with Hadoop, NoSQL and other relational database sources. Special Integration to R. Connectivity to 3 rd party solutions. include Cloudera Hadoop distribution and management software
Teradata	Aster Database 5.0 with SQL-MapReduce Teradata Database 14 Hadoop Integrator	Aster MapReduce Appliance Active Enterprise Data Warehouse Extreme Performance Appliance Data Warehouse Appliance Extreme Data Appliance Data Mart Appliance	Data Integration with Hadoop, NoSQL and other relational database sources. Special Integration to SAS. Connectivity to 3rd party solutions.
Cloudera	Hadoop platform distribution (CDH) Data Integrator Automated Cluster Management Search, explore, visualize and analyze engines Web applications that enable you to interact with a CDH cluster(HUE)	CDH (Cloudera’s Distribution including Apache Hadoop) Cloudera Express Cloudera Ent	Data Integration with Hadoop, NoSQL and other relational database sources. Connectivity to 3rd party solutions. Connectors for Netezza, Teradata,Tableu, Microstrategy

C. Big Data Commercial Evaluation

This evaluation had in consideration most of the important features in Big Data:

Data Integration: Allow a quick and easy integration of multiple data sources (unstructured, semi-structured and structured). At same time it allows organizations to execute data analytics tasks with data virtualization.

Data Visualization: Is the way of the information (knowledge) is displayed and make available. Solutions with this features allow not only the data preparation but also has a data visualization shape in order to present the information stored.

Big Data Analytics: Allow to analyze large data sets containing a variety of data types. Enables organizations to analyze a mix of structured, semi-structured and unstructured data in search of new knowledge and valuable business information.

Interactive Search: Allow an easy way to execute queries by searching particular data / information that the "user" wants (making for example data comparison and if available drill-down and roll-up). Sometimes this type of features has associated information retrieval algorithms.

Text Analytics: Capability to transforms free-form text documents into a chosen intermediate form and deduces patterns on knowledge from the intermediate form.

Real Time: This is a required feature to the Intelligent Systems [9]. It means that a system is able to execute the main tasks in real-time in order to support Data Processing, Data Integration, Data Mining and Data Visualization.

Batch Processing: Also known as intelligent agents [10, 11] it is the execution of a set of tasks (e.g. data processing) on a system without manual intervention.

Table 3 - Big Vendor Solutions - Evaluation

Features	IBM	Oracle	EMC	Teradata	Cloudera
Data Integration	✓	✓	✓	✓	✓
Data Visualization	✓		✓	✓	
Big Data Analytics	✓			✓	
Interactive Search	✓	✓	✓	✓	
Text Analytics	✓				
Real-Time	✓	✓	✓	✓	✓
Batch processing	✓	✓	✓	✓	✓

After analyzing table 3 is easy to understand that there are three essential features for all vendors: Data Integration, Real-Time and Batch processing. From the five vendors evaluated only the IBM has all the features presented in their solution. Following the features list appears Teradata and EMC. The fact of IBM arises in first place does not means that IBM is the best solution (the decision-makers should to choose their solution according to project requirements). Actually it is possible combining solutions in order to achieve a better result. For example if the decision-maker prefers working with Teradata can combine it with a tool presented in table 1 / figure 3 in order to have the same features than IBM or even better.

V. CONCLUSION

Being real-time and batch processing two required features to deploy an Intelligent System or application all the vendors are able to support this deployment. However the vendor should be chosen in accordance to the project requirements. After choose a Big Data architecture, they can choose an Analytic Open Source Tool in order to support the data analysis process.

Table 2 helps the decision-makers to have a better understanding of which vendor is more suitable for their project. This study allows to have a better understanding about which are the main features of each vendor / solution.

One of the main goals of this paper is to avoid the researcher effort in searching for Big Data solutions to support their system. By reading this paper they can know what are the main open-source technologies and vendors. They also can know what the key vendors offer and what is the hardware appliance and the connectors provided by each.

In the future two studies: Commercial vs Open Source and Appliance vs Cluster will be performed with the goal to display the pros and cons of each solution. Then a Big Data architecture will be designed taken in considerations the studies results.

ACKNOWLEDGMENTS

The authors would like to thank Deloitte for some of the resources provided and FCT (Foundation of Science and Technology, Portugal) for the financial support through the contract PTDC/EEI-SII/1302/2012 (INTCare II). This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013.

REFERENCES

- [1] J. P. Dijcks, "Oracle: Big data for the enterprise," *Oracle White Paper*, 2012.
- [2] M. M. Gobble, "Big data: The next big thing in innovation," *Research-Technology Management*, vol. 56, p. 64, 2013.
- [3] P. Sridhar and N. Dharmaji, "A comparative study on how big data is scaling business intelligence and analytics," *Int. J. Enhanced Res. Sci. Technol. Eng.*, vol. 2, pp. 87-96, 2013.
- [4] G. Hiong, "Open Source and Commercial Software: An In-depth Analysis of the Issues," ed, 2004.
- [5] Deloitte, "Title," unpublished.
- [6] P. Pääkkönen and D. Pakkala, "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems," *Big Data Research*, 2015.
- [7] I. I. Center. (2014, April, 24, 2015). *Big Data in the Cloud: Converging Technologies*. Available: <http://www.bigdata-startups.com/open-source-tools/%2012-08-2014/>
- [8] N. Hague. (2014, April, 24, 2015). *BigData Startups - the big data knowledge platform*. Available: <http://www.bigdata-startups.com/open-source-tools/%2012-08-2014/>
- [9] F. Portela, M. F. Santos, P. Gago, Á. Silva, F. Rua, A. Abelha, et al., "Enabling real-time intelligent decision support in intensive care," in *25th European Simulation and Modelling Conference- ESM2011*, Guimarães, Portugal, 2011, p. 446 pages.
- [10] L. Cardoso, F. Marins, F. Portela, M. Santos, A. Abelha, and J. Machado, "The Next Generation of Interoperability Agents in Healthcare," *International journal of environmental research and public health*, vol. 11, pp. 5349-5371, 2014.
- [11] M. Wooldridge, "Intelligent agents," in *Multiagent systems: a modern approach to distributed artificial intelligence*, ed: MIT Press, 1999, pp. 27-77.

Proposed Runtime Decision Making framework for Autonomic Software Systems

Sandeep Kumar Chauhan, Arun Sharma, P S Grover

Abstract: The core feature of Autonomic Software Systems is its self-healing and self-learning behaviour. Autonomic Software systems are capable of managing their behaviour and resources automatically to accomplish a given goal, despite changes in the environmental conditions. The responses to the environmental changes are achieved by self healing. The self-healing and learning can be achieved using runtime decision making in Autonomic Oriented Production systems. There is a great need for a robust framework for developing decision making frameworks for Autonomic software systems. In this paper we proposed a generic framework for these systems, based on system's previous knowledge and learning.

Keywords: Autonomic System, Self-Healing, Intelligent techniques, Framework for Autonomic

I. INTRODUCTION

Large IT organisations are facing a challenge due to high rate of growth in IT systems and applications. Due to this, a high percentage of their budget goes to manage these heterogeneous and complex IT systems. There is a strong need for such systems which can manage itself against the unforeseen problems to cut down the operational cost. The self-management [1] feature of Autonomic Computing [2] aims to manage (fault detection, diagnose, and repair) complex IT environment at a lower cost to help large organisations. Development, operation and maintenance of self-managing systems are extremely challenging. Autonomic computing is still evolving and proven framework and standards for developing autonomic system is still a research area. Systems designed to be self-healing [3] should be able to heal themselves at runtime in response to changing environmental or operational circumstances. There can be several methods to achieve the self-healing feature in the distributed and complex IT system. In absence of a generic framework, applications or systems are customized as per need without any standardization. Therefore, there is great need for a generic and standard framework for developing and managing self healing Autonomic.

Sandeep Chauhan is research scholar in Mewar University, Chittorgarh, India (email: chauhansandeep@hotmail.com)

Arun Sharma is working as Associate Professor- IT at Indira Gandhi Delhi Technical University for Women, Delhi India. (email: arunsharma2303@gmail.com)

P S Grover has been Director, Computer Science at Delhi University, Delhi India and presently associated with KIIT, Gurgaon, India. (email: drpsgrover@gmail.com)

The standard framework will help future implementation to be done using component based model where the autonomic components can be chosen from various vendors and also help IT organisations to have plug and play feature to manage their distributed and complex IT infrastructure. In this paper we propose a generic framework for developing self healing Autonomic systems. This framework may be used to develop future autonomic systems for complex IT software and may be based on cognitive learning and other soft computing techniques.

II. RELATED WORK

There was study done by Darius Plikyans on self-organization system and suggested a model OSIMAS[21] (oscillation-based multi-agent system) based on self-organization behavior. The vision for self healing in autonomic computing was the basis for several research activities in the past years in both industry and academia. Many approaches extend the traditional software design, proposing component-based models with dynamic configuration or autonomic capabilities.

A. Unity Project

Unity is a research project, carried out at IBM's Thomas J. Watson Research Center, that explored some of the behaviours and relationships that allowed complex computing systems to self-manage [10] having its four basic self-CHOP (configure, healing, optimise and protecting) properties. The four principal aspects examined with the Unity prototype environment are (i) the overall architecture of the system, (ii) the role of utility functions in decision making within the system, (iii) the way the system uses the goal-driven assembly to self-configure, and (iv) the design patterns that enable self-healing within the system. The IBM Autonomic Computing Toolkit is designed for users who want to learn, adapt, and develop autonomic behaviour in their products and systems using recommended tools, technologies, and scenarios. The autonomic tools and scenarios can be grouped into three main categories: (i) problem determination, (ii) solution installation and deployment, and (iii) integrated solutions [11].

B. KX Project

The goal of the KX Project, led by Kaiser [12], was to inject autonomic computing technology into legacy software systems without any need to understand or

modify the code of the existing system. The project designed a meta-architecture implemented as an active middleware infrastructure to add autonomic services explicitly via an attached feedback loop that provided continuous monitoring along with a capability to reconfiguration or repair if required. The lightweight design and separation of concerns enable easy adoption of the entire infrastructure and also allows the consumption of an individual component.

C. *Rainbow Project*

The Rainbow Project, led by Garlan, investigated the use of software architectural models at runtime as the basis for reflection and dynamic adaptation [13], [14]. The project aimed to provide reduced user intervention while adopting new the configuration /system changes to make the overall system up and running to meet the quality goals. Project also aimed to improve the dependability of changes, and support a new breed of systems that can modify itself in response to dynamic changes in the environment. There was another sub project DiscoTect was developed to produce architectural views by observing the running system [15]. This helped in overall analysis of Rainbow project.

D. *AutoMate Project*

Project AutoMate investigated autonomic solutions that are inspired by biological systems and deals with similar challenges of highly complex system that are dynamic and heterogeneous in nature. Project AutoMate aimed at developing conceptual models and implementation architectures that can enable the development and execution of such self-managing Grid applications. The overall objective of the AutoMate Project is to investigate key technologies to enable the development of autonomic grid applications that are context aware and capable of basic autonomic properties [16], [17]. The project investigated the programming models and frameworks used in autonomic applications. It implemented autonomic component management comprising CHOP properties to optimize resource utilization and application performance in situations where computational characteristics and/or resource characteristics may change.

III. SELF HEALING USING RUNTIME DECISION MAKER

A blueprint for self-managing systems has been proposed in 'The Vision of Autonomic Computing' [1]. A self-healing system is split into an Autonomic Manager and a Managed System [2]. The autonomic manager utilises the self-healing logic by implementing a MAPE-K(Monitor /Analyze /Plan /Execute and Knowledge) [1] feedback loop that controls the managed system through Sensors and Effectors. Sensors are subject to perform continuous monitoring. They accept input from the environment through managed resource touch points. It provides its output to the monitor routine that can

collect information for proceeding to other levels. Effectors are self-adjustors. They provide appropriate response corresponding to the sensor state and thus the knowledge acquired.

Other relevant work includes the DARPA-funded Dynamic Assembly for System Adaptability, Dependability, and Assurance (DASADA) initiative [8]. Here the focus is on software engineering concerns including; a generative programming model for finer-grained dynamic and predictable software adaptation. Using an architecture-driven approach incorporating exploration and estimation enables the software to interact with the executing system and collect raw measurement data for translation into suitable metric information for error recovery through adaptation.

The MAPE-K blueprint is the basic principle of our framework that utilise the runtime decision maker in conjunction.

In today's complex and distributed IT systems, availability has become top priority aspect and autonomic computing using self-healing helps to maintain availability at lowers costs. Apart from availability in complete solution. Each organisation wants to adopt ACME (Availability, Changeability, Maintainability and Evolutionary Growth) at low cost. To achieve ACME for self healing systems; it is very important to detect the root cause of the failure so that an appropriate corrective action can be taken to fix the problem. The quicker and correct identification of the issue will help the autonomic system to recover from that situation. The traditional approaches for error detections [4] can be used identify the cause which may fall under one of the category such as Syntactic faults, Semantic faults, Service faults, Communication / interaction faults and Exceptions [5]. Traditionally based on the fault outcome a manual intervention is required. In many cases the fix may require a code/configuration change which will follow a very tedious process of validation and verification before it can be implemented in the Production systems. To make systems highly available the runtime decision maker can help systems to get the required resolution in Production system itself. The resolution may be a configuration change or can be an auto generated code which can fix the issue.

The runtime decision maker [4] can identify a suitable resolution to be implemented during runtime to heal the system from internal and external failures. Runtime decision maker [4] will make use of pattern training, cognitive learning and artificial intelligence to correctly identify a resolution based on the past experience.

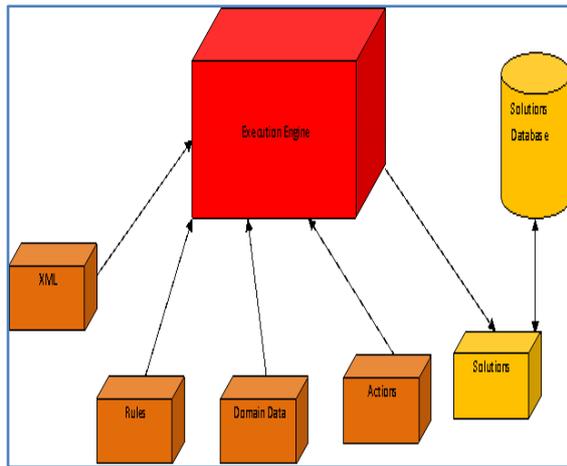


Figure 1: High Level view of Decision Maker [4]

Runtime decision maker makes use of pre-configured rules, actions, domain data and suggested solution to implement the correct solution. The validity and correctness of the implemented solution will always depend upon various aspect of system implementation such as decision tree templates, knowledge based discovery and correct fuzzy logic implementation for processing the actions.

IV. PROPOSED SELF-HEALING FRAMEWORK

The self-healing framework architecture is based on MAPE-K [1] [2], a recommendation, which is a collection of abstract components. There will be two parts of the application

- (i) the base application – Managed System
- (ii) the autonomic components for monitoring and self-healing – Autonomic Manager

The proposed framework would be a combination of monitoring, detecting, recovery service along with decision execution engine based on rules and their actions. The following components are the recommendation parts of the self-healing framework:

A. Monitoring Agent:

The main job of Monitoring Agent (MA) will be to monitor the base/observed software system for any faults or abnormal behaviour. The proposed monitoring agent feature can be implemented based on Kieker Framework [8] recommendation that will allow continuous monitoring with very little overheads on the production system. The only restriction of this product is that it is restricted to Java objects but support for other environment such as .NET is under development. The MA component will facilitate to monitor and analyze the runtime behaviour of software systems. MA will monitor the behaviour, based on a pre-defined and configurable monitoring rule.

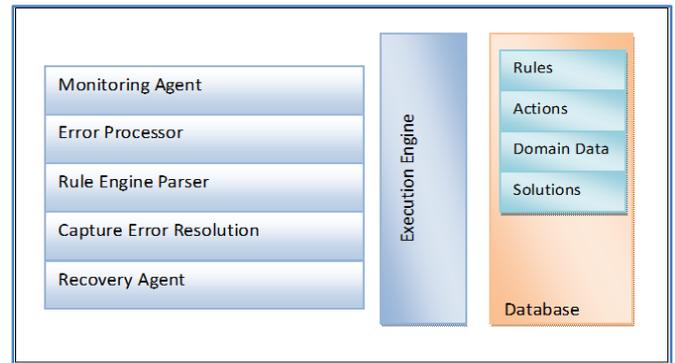


Figure 2: Framework for self-healing systems

V. ERROR PROCESSOR

Error Processor (EP) will receive an input from MA during runtime. Based on the input, EP will process the error messages into a standard XML format to standardise the error format from various input systems. EP will use configurable information to generate a proper error message which can be understood by the decision engine component of runtime decision maker [4]. The main job of the component will be to format and enrich the error message/code received from MA. The enrichment rules can read from an external database/property file to cater vast variety of systems. Error processor will be adding some more Meta information such as node name, timestamp and demographics information in case of distributed systems to the actual error.

VI. RULE ENGINE PARSER

Rule Engine (RE) Parser component will be responsible for parsing the rules stored in the system. These rules can be associated with monitoring, enrichment of error message or related to self-healing action. RE will be based on the rule parser configuration. While implementing a system that needs complex rule parsing only the parser component may need to be developed so that RE component can parse new set of rules to process the required information. RE will get most of its knowledge from the previously stored information in database. Each rule will have many to one mapping with Action which are also stored in the DB. There are number of existing rule engine exists such as Java Rule Engine API or any existing Business Process Execution Language (BPEL) can be customised for this kind of implementation. Soft Computing techniques like Fuzzy Logic, Artificial Neural Network may also be used to identify the best decision. An integrated bio-system modeling [22] can be referred in building a robust model which help handling complex situations that deals with greater uncertainties. Another framework on fuzzy system has been recommend in paper 'A Fuzzy Systems Framework for Solving Real World Problems'[23] is very good reference implementing rule engine parser module. The cased based reasoning can be useful for solving many complex problem in real world IT

problems and it can be the basis for developing the overall system for rule engine parser to have robust autonomic system.

VII. CAPTURE ERROR RESOLUTION:

Capture Error (CE) Resolution component will be responsible for cognitive learning. This module will check of existing resolution so that an action can be taken to resolve the issue. The cases where desired rules and actions are not available and system needs human intervention for the correct resolution, the CE component will capture and store the rule and its action in database for future usage. This module will have complex mathematical implementation to match the exact solution for a given problem. For simple cases that can be resolved with IF THEN ELSE, simple API is suffice. This implementation will be based on natural language processing. In cases where engineer can update the system solution in natural language system should be to interpret it and work accordingly. The concept of Hierarchical temporal memory [18] can implement to build such system. The temporal memory based system can be trained first to accept and understand user commands and during recognition phase it will find out the best matching information so that based on that a correct action can be taken to resolve the error.

VIII. RECOVERY AGENT

Recovery Agent (AG) component will be responsible for downloading, installing, deployment and testing the fix before adding the change in production environment. This component/module will take inputs from monitoring capture error resolution. Based on the resolution state, this module will act on getting the correct components in the system. This will be analogous to Effector [2].

IX. EXECUTION ENGINE

Execution Engine (EE) component will work as orchestrator for all other components mentioned for the self healing framework. This component will decide the flow of all the components in the system. It will be responsible to creating the action job queue that will be initiated by recovery agent module. This will keep track of all the running /completed/resumed/suspended actions. It should be able to help administrators to see the current action status so that they can take manual action if required in some special cases.

X. ISSUES AND CHALLENGES

Framework and designing of Autonomic Systems is a complex task and still a research area. The implementation of run time Decision Maker for Autonomic System can achieved using various Artificial Intelligence techniques. However, there are few issues and challenges:

- a) The RDM should be able to install and implement for unknown architecture and domains. For known system, the system will refer to existing rules, in case of new error scenario the system may not be trained enough to take the right decision. Also until the machine learning and intelligence mature the organisation may not take a risk to implement such system in Production environment.
- b) The cognitive learning [19] and natural language processing are still in its early stages and RDM will require a robust system that will learn and apply rules in dynamic complex IT environment.
- c) Another big challenge will be testing (validation/verification) [20] of RDM systems. It will be a complex task to create the testing environment and establish coordination among various components of Autonomic system that is using RDM system. The existing testing automation can execute test cases based on previous test cases and data. In case of run time implementation, system will require new test cases that can test the system thoroughly. This also require a mature technology implementation so that all test cases and their respective data can be generated in run time as well to make the system more robust and ready to be deployed/install in production systems.
- d) Still researchers are working on the Autonomic component standards. The un-availability of standards result into complex and difficult the component interrelationship.
- e) Achieving the optimisation and protection in running system is still a very difficult task. Also the run time decision maker need to make sure that code is not using a boiler plate which makes a lot of duplication for implementation. This can defeat the objective of optimisation. Also in the case of protection system can implement heuristic techniques but they will be based previous experience and may not help to make system fully protected from future errors completely.

XI. CONCLUSION

As we know that we are experiencing a huge exponential growth in IT systems both at servers and commodity machines. This has resulted into a complex IT infrastructure. In recent past the computing trend has shifted to commodity hardware and peer to peer processing but still for large IT organisation to keep the system up and running is still a challenge. We believe that introduction and implementation of run time decision maker framework may help the future IT system to be more robust and available to end user 99.99%. There are still open challenges in implementation a very robust artificial intelligent system. There is lot of research is being done in Artificial intelligence domain and this will surely help to achieve all the goals of run time decision make framework.

The realisation of the autonomic vision can be possible for industry and academia by leveraging the suggested framework by implementing it using widely used programming paradigm and tools. We are sure that in near future we will have a strong implementation available for this framework which will help IT organisations to operate on lesser costs with high availability.

REFERENCES

- [1] Kephart J. O., Chess D. M., "The Vision of Autonomic Computing" *Computer, IEEE*, Vol. 36, Iss. 1, January 2003, pp: 41-50.
- [2] IBM, "An Architectural Blueprint for autonomic computing", *IBM White Paper*, Vol. 7, June 2005 pp: 1-31.
- [3] Julie A. McCann, Markus H., "Evaluation issues in Autonomic Computing, Grid and Cooperative Computing – GCC 2004 Workshops, pp. 597-608.
- [4] Chauhan S. K., Sharma A., "Runtime Decision Making for Developing Autonomic Systems", *International Journal of Computing, Intelligent and Communication Technologies*, Vol., Iss. 2, May 2013.
- [5] Michael J., Jing Z., David R., and John S., "A modelling framework for self-healing software systems", *10th International Conference on model Driven Engineering Languages and Systems*, 2007.
- [6] Ehlers J., and Wilhelm H., "A self-adaptive monitoring framework for component-based software systems." *Software Architecture*, Springer Berlin Heidelberg, 2011, pp: 278-286.
- [7] Salehie, M., and Ladan T., "Self-adaptive software: Landscape and research challenges", *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* Vol. 4, Iss. 2 2009, pp: 1-14.
- [8] Van Hoorn A., Rohr M., Hasselbring W., J. Waller, J. Ehlers, S. Frey, D. Kieselhorst. "Continuous monitoring of software services: Design and application of the Kieker framework." Dept. of Computer Science, University of Kiel, 2009.
- [9] DARPA, Dynamic Assembly for Systems Adaptability, Dependability, and Assurance (DASADA), <http://www.rl.af.mil/tech/programs/dasada/>, 2000.
- [10] Chess D. M., Segal A., Whalley I. & White S. R., "Unity: Experiences with a Prototype Autonomic Computing System," *Proceedings of the IEEE First International Conference on Autonomic Computing (ICAC 2004)*. New York, NY, May 17-18, 2004, Los Alamitos, CA: IEEE Computer Society, 2004, pp: 140-147.
- [11] IBM Corporation "Autonomic Computing Toolkit.", <http://www-128.ibm.com/developerworks/autonomic/overview.html> (2005).
- [12] Kaiser, Gail; Parekh, Janak; Gross, Philip; & Valetto, Giuseppe," Kinaesthetic eXtreme: An External Infrastructure for Monitoring Distributed Legacy Systems", *Proceedings of the Autonomic Computing Workshop Fifth Annual International Workshop on Active Middleware Services (AMS 2003)*, Seattle, WA, June 25, 2003. Los Alamitos, CA: IEEE Computer Society, 2003, pp: 22-31.
- [13] Cheng S., Huang, An-Cheng, Garland D., Schmerl B., & Steenkiste P., "Rainbow: Architecture-Based Self Adaptation with Reusable Infrastructure", *IEEE Computer*, Vol. 37, Iss. 10, October 2004, pp: 46-54.
- [14] Garland, David & Schmerl, Bradley. "Using Architectural Models at Runtime: Research Challenges," *Proceedings of the First European Workshop on Software Architecture (EWSA 2004)*, St. Andrews, Scotland, Berlin, Germany: Springer-Verlag, 2004, pp: 200-205.
- [15] Hong Y, Garland D, Schmerl B., Aldrich, J. & Kazman R., "DiscoTect: A System for Discovering Architectures from Running Systems" *Proceedings of the ACM/IEEE 26th International Conference on Software Engineering (ICSE 2004)*, Edinburgh, Scotland, May 23-28, 2004. Los Alamitos, CA: IEEE Computer Society, 2004, pp: 470-479.
- [16] Parashar, Manish, "Project AutoMate" <http://automate.rutgers.edu/> (2003).
- [17] Agarwal, M., Bhat, V., Liu, H., Matossian, V., Putty, V., Schmidt, C., Zhang, G., Zhen, L., Parashar, M., Rutgers, Khargharia, B., & Hariri, S., "AutoMate: Enabling Autonomic Applications on the Grid," *Proceedings of the Autonomic Computing Workshop Fifth Annual International Workshop on Active Middleware Services (AMS 2003)*. Seattle, WA, June 25, 2003. Los Alamitos, CA: IEEE Computer Society, 2003, pp: 48-59.
- [18] Jeff Hawkins and Dileep George, "Hierarchical Temporal Memory - Concepts, Theory, and Terminology", Numenta Inc., 2006.
- [19] Dale Shaffer Wendy Doube Juhani Tuovinen , "Applying Cognitive Load Theory to Computer Science Education", *Proc. Joint Conf. EASE & PPIG 2003*
- [20] Arun Sharma, Sandeep Chauhan and P.S Grover, "Autonomic computing : Paradigm Shift for Software Development" , CSI Communications, Vol 35, Iss. No 5, Aug 2011.
- [21] Darius Plikynas, "Towards Representation of Agents and Social Systems Using Field-Theoretical Approach", *WSEAS Transactions on Systems*, Vol. 13, 2014, pp. 730-744
- [22] S. Vassileva, "Advanced Fuzzy Modeling of Integrated Bio-systems", *WSEAS transactions on Systems*, Iss. 7, Vol. 11, 2012, pp. 234-243.
- [23] Michael Gr. Voskoglou, "A Fuzzy Systems Framework for Solving Real World Problems", *WSEAS transactions on Systems*, Iss. 8, Vol. 9, August 2010, pp. 875-884

Modular system for gathering and classification of SIP attacks

J. Safarik, M. Voznak, J. Slachta, L. Macura, F. Rezac and J. Rozhon

Abstract— The article deals with an application of artificial intelligence on classification of attacks in IP telephony from data on honeypots. Data about attacks on these honeypots are collected on a centralized server and then classified in the neural network. The paper describes inner structure of used neural networks and also information about their implementation. The trained neural network is capable to classify the most common used VoIP attacks. Two approaches are investigated in this paper, the first one is based on MLP (Multilayer Perceptrons) and the second one on SOM (Self-organizing Maps). With the proposed methods is possible to detect malicious behavior in a different part of networks and achieved results confirm applicability of the implemented approaches.

Keywords— SIP attacks, classification, multilayer perceptron network, self-organizing maps, neural networks.

I. INTRODUCTION

THE main purpose of a honeypot is to simulate the real system and interact with anyone in the same way as the production system would. It watches the behavior of anyone who interacts with it [1]. Monitoring of VoIP infrastructure, IDS/IPS (Intrusion Detection/Prevention Systems) or honeypots application can detect these attacks and malicious activity in the network [2], [3], [4]. Some of these mechanisms can disrupt or mitigate certain types of attacks, but there is still much of remaining attacks, which can impact VoIP servers. The information about SIP attacks from a honeypot application brings valuable source of network attacks. However analysis of data from these honeypots, especially in large or divergent network, cause unwanted overhead for network administrators.

This paper is organized as follows. In the chapter VoIP Honeypot we provide a close look on individual honeypot applications and its integration in honeypot cloud solution. In

This research has been supported by the Ministry of Education of the Czech Republic within the project LM2010005.

J. Safarik, J. Slachta, F. Rezac and J. Rozhon are PhD. students with Dept. of Telecommunications, Technical University of Ostrava and also researchers with Dept. of Multimedia in CESNET, Zikova 4, 160 00 Prague 6, Czech Republic (e-mail: safarik@cesnet.cz, slachta@cesnet.cz, filip@cesnet.cz, rozhon@cesnet.cz).

L. Macura is a network administrator of Silesian University in Opava and he is also a researcher with Dept. of Multimedia in CESNET, Zikova 4, 160 00 Prague 6, Czech Republic (e-mail: macura@cesnet.cz).

M. Voznak is an Associate Professor with Dept. of Telecommunications, VSB-Technical University of Ostrava (17. listopadu 15, 708 33 Ostrava, Czech Rep.) and he is also a researcher with Dept. of Multimedia in CESNET (Zikova 4, 160 00 Prague 6, Czech Rep.), corresponding author provides phone: +420-603565965; e-mail: voznak@ieee.org.

the next chapter we propose our solution collecting data from honeypots. The core of the presented research is in following chapters where we discuss design of classifiers based on the artificial neural networks and achieved results.

II. HONEYPOTS

Running individual honeypot application on a single server brings valuable information. Exceeding numbers of running honeypots, especially if running in different networks on various geographic locations, causes unwanted overhead in data analysis. Without an automatic aggregation mechanism, this situation leads to decreasing profit from honeypot's data.

In previous work [5], we created a testing topology to measure DoS effectiveness and for further testing. It consisted of SIP proxy server, hacker's PC and some endpoint devices. The Sipp programme was primarily used to simulate calls and to carry out SIP proxy stress tests. The application SIPP is an open source traffic generator that was designed specifically for testing purposes. Sipp is capable of simulation of both UAC and UAS and can also generate both signaling and media traffic. The DoS attack scenario with selected SIP methods was applied and from the load on server was obvious that the SIP server was affected by the DoS attack and the most efficient attack is based on method REGISTER and INVITE due to increasing computational complexity caused by the required authentication. As we acquired practical experience with threats in IP telephony [6], we decided to examine honeypots and their usability for detection of attacks [7], [8].

Honeypot applications log and monitor malicious activity. Nowadays exists many honeypot solutions emulating both single and multiple services. Because of our focus on IP telephony we decide to deploy suitable honeypots. Each honeypot emulate different aspects of network, its features are described below.

A. Artemisa

An Artemisa honeypot can be deployed in any VoIP infrastructure which uses a SIP protocol. The programme connects to SIP proxy with the extensions defined in a configuration file. The extensions should be within the range which is typically used for real accounts. The main purpose is to establish a better masking against the potential attackers. Artemisa itself does not simulate PBX but rather an active endpoint device.

B. Kippo

The second mentioned honeypot is based on different foundations. It is not VoIP oriented as Artemisa. It simulates a SSH server. When someone tries to connect to a server with a honeypot running on it, the `twistd` application redirects this user to the honeypot. This happens where the user IP address is not included in the list of permitted IP address. Once the connection with the honeypot is established, the attacker must enter correct username and password. Kippo logs every login attempt. Where the entered combination is valid, the intruder is granted access to a fake filesystem. Every command entered into the honeypot is logged and behaviour typical for a particular command is emulated (for the most common commands only). If the user tries to download something from the Internet, Kippo saves this file into a secure folder for further examination. All logs made by Kippo are saved in a MySQL database which facilitates the subsequent analysis.

C. Dionaean

Two previously mentioned honeypots were single service oriented ones. Dionaean belongs to a multi-service oriented honeypot which can simulate many services at a time. Typically are information from these multiple services only general but Dionaean serves only small number of them like SMB (Microsoft's printers, files, serial ports sharing protocol), HTTP, FTP, TFTP, MSSQL (Microsoft SQL server), SIP protocols. Attackers abuse these protocols in most cases. Dionaean has also ability to save malicious content needed by hackers securely, but as a contrary to Kippo can also emulate code from these files.

Describing features of all this protocols is beyond the scope of this paper and further features focus only on the SIP protocol. Dionaean works in a different way as Artemisa. There is no need for connecting to an external (or production) VoIP server. It simply waits for any SIP message and tries to answer it. It supports all SIP requests from RFC 3261 (REGISTER, INVITE, ACK, CANCEL, BYE, OPTIONS). Dionaean supports multiple SIP sessions and RTP audio streams (data from stream can be recorded). For better simulation of a real IP telephony system, it is possible to configure different user agent phone mimics with custom username, password combinations. There is functionality for a different pickup time on simulated phones via pickup delay feature. All traffic is monitored, and logs are saved in plain-text files and in sqlite database.

D. Kojoney

The Kojoney is a honeypot written in Python, licensed under GNU GPL and is not oriented directly focused on VoIP, this honeypot simulates the SSH server. Attackers are redirected to the honeypot and do not have possibility to communicate with a real system. The Kojoney is usually configured with many number of accounts and passwords increasing successful authentication of attackers. As attackers are logged on they enter commands and the honeypot is able to response with a predefined texts which can be easily configured in the

Kojoney's text module. All activity is saved in log files of the honeypot. These reports contain information on successful and unsuccessful login, all commands submitted within particular session and source IP addresses of attackers.

III. DESIGN AND IMPLEMENTATION OF HONEYPOT CLOUD

In order to collect data from honeypots we designed and implemented solution which includes two parts – client and server [9].

A. Client Part

Client part of our honeypot image consists of several components, the crucial part is a honeypot application. Based on our previous experience [7] we decided to include only honeypot Dionaean in final image. The structure of client side is illustrated in Fig. 1.

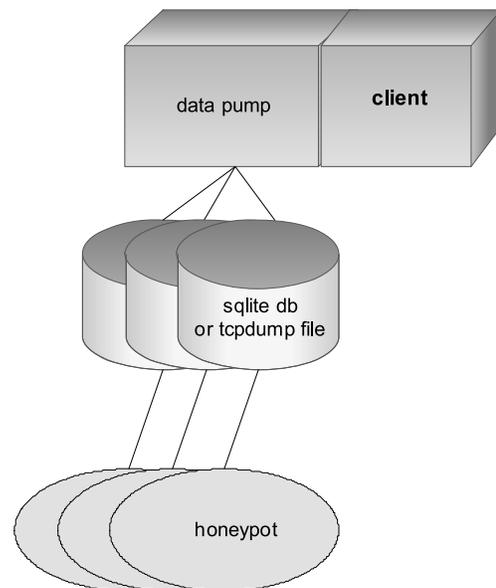


Fig. 1 Honeypot concept of client side.

The honeypot supports logging information into a database, nevertheless we are collect information on honeypot directly from `tcpdump` as well. There is a data pump for data mining and information are gathered from these databases. Communication with server handles another part of application which we developed for honeypot image.

Whole communication between client and server application is encrypted. This feature also brings a possibility to easily authenticate client nodes with certificates. Data gathered through data pump are converted into serialized objects. These objects are periodically sent to server for further operations and analysis.

Our honeypot image is under intensive testing, on several universities and in home environments of employees as well, and practically ready for final deployment. There is pre-configured firewall allowing only traffic specific for each honeypot application and internal client communication with server.

B. Server Structure

Some features of server-side applications were already described in the previous section. Server's main task is collection of received data and their analysis. Due to this functionality is whole application implemented as a ROLAP (Relational online analytical processing) and the concept is depicted on Fig. 2. Otherwise from a client (honeypot image), server is considered as enterprise application and not as an image. Before ROLAP receives data from nodes and stores the data to database, several last steps are performed. Nodes send already cleaned and fixed data structures, so ROLAP can proceed to data transformation and integration according to data store scheme. All data are divided in two groups - on the dimensions (date and time, honeypot, locations, etc) and the facts. Data warehouse consists of a star or a constellation design. Internally run data store on the relational database.

Built-in aggregation function ensure the aggregation of all stored data in groups by hours, days, months and years. This function is implemented for higher performance of database. The server also comprises a self-monitoring mechanism for logging important information like system logins and analysis of actions or failures. System log contains details information about each honeypot node outage.

Each honeypot node must be activated and authorized before establishing a full connection between server and client. Node authentication is included in SSH tunnel assembly, and there is no other mechanism. After successful authentication and authorization start client periodical notification and data transfers. The gathered data from honeypot node are sent periodically each day. Server operates with data which are only one day old in the worst case.

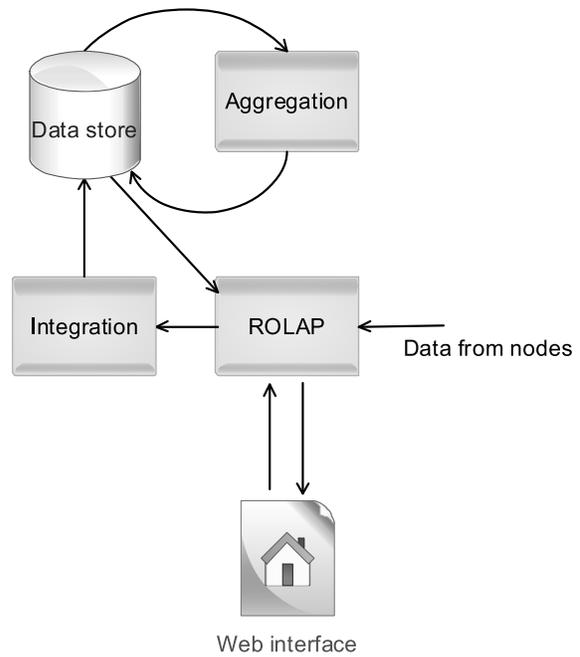


Fig. 2 Server side application structure.

It is possible to change this interval to a minimum period of an hour. System contains a deactivation function for case of removing honeypot.

Each attack detected on any honeypot is checked for source location. This feature works via IP address localization. System automatically downloads lists of IP address range for all states. Final file runs through parsing function and the system operates with information from RIR (regional Internet registry) databases only one day old.



Fig. 3 Data analysis of collected data from honeypots.

Using IP address localization brings another information value to already collected information. Whole system is

controlled via built-in web user interface running on secured HTTPS protocol. Access to the system is protected with user accounts, in the future is planned implementation of other login options using Shibboleth or OpenID.

A wireframe of web user interface, after user login, is depicted in Fig. 3. The design of the interface is involved as an intuitive for users enabling various data analysis. Last part of server side architecture is a VoIP attack classification engine which is still under development and we present our achieved results in next chapter.

IV. THE PROPOSED CLASSIFICATION ENGINE

The system consists of the server and distributed detector nodes. The architecture was described in a previous paper by the author [9] and [10]. All attack data from nodes is sent directly to the server, after successful authentication, server parses uploaded files and starts attack classification with an artificial neural network (ANN).

The latest improvement on the server side is authentication and authorization handling, responsive design of user interface and sliding window correction for data aggregation. Each node simply cleans data from source application and upload them once per hour. The server is responsible for the aggregation. Therefore, it needs to merge attacks which prevail in two or more upload sets. Those attacks are related and must be classified together. The server aggregate attacks to groups by source IP addresses. Firstly check, if it receives the following packet from the same IP address. Used port or transport protocol can vary between connections, important is timestamp of each SIP message. If the message is delivered in less than 5 minutes after the last message from a single source, it will belong to same aggregated group as the previous message. If the time extends 5 minutes, server creates new aggregated group.

Both implemented ANNs mentioned in following sub-chapters use input structure consisting of 10 parameters (also called input vector): number of connections in attack, all RFC 3261 SIP messages count, subscribe messages count, total SIP messages per connections ratio, attack duration. Final classification classes follow only real attacks detected on Dionaea honeypot. Those are call testing, registration attempts, DoS attacks, scanning, information gathering and fuzzing.

A. Classifier based on MLP

The classification engine uses multilayer perceptron network (MLP). MLP models the neural system of mammals with millions of interconnected cells in a complex arrangement. The MLP used for classification works with four layers of neurons as is depicted in Fig. 4, each neuron in one layer is interconnected with all neurons in the upper layer. We distribute the number of neurons in layers as 10-30-24-8.

Because MLP belongs to supervised type of neural network, we had to prepare a training set. This set contains only attack representatives detected with on distributed node, not representatives of all possible SIP attacks. This approach ensures that MLP recognizes only current and relevant attacks.

If a new kind of attack appears, we must train MLP with new training set.

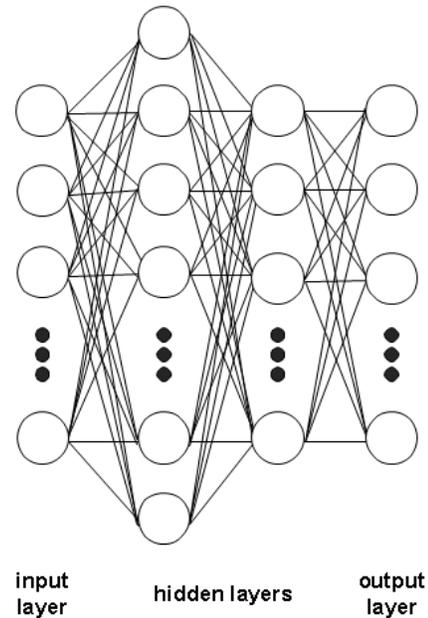


Fig. 4 MLP neural network topology

The MLP is trained by a backpropagation algorithm, which adapts weights of interconnections between neurons. The current training set contain 104 representatives, 13 for each class of attack. Whole backpropagation algorithm and MLP architecture are described in [9]. Principles of neural networks are very well described in [11] and introduction to their implementation in JAVA is included in [12].

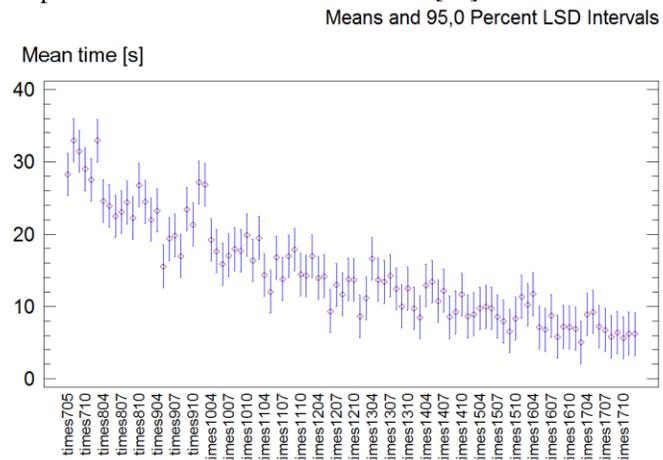


Fig. 5 Mean times of 100 backpropagation iteration with different inner structure.

The performed tests of our implemented approach proved different mean times of learning for different inner structures, see Fig. 5 (timesXXXX – XX means number of neurons in first hidden layer, YY – number of neurons in second hidden layer). The impact of the structure is evident only for backpropagation learning time. With higher numbers of neurons in hidden layers, the mean time of backpropagation

learning decreases. On the other hand, raise memory and computational requirements of MLP neural network.

After proper learning, inner structure has no more statistically significant impact on attack classification. The final neural network structure for new generation network contains 30 and 24 neurons in hidden layers, because of conducted investigations and tests. MLP network is evaluated as learned, if there correctly identify more than 95% of items in the training set (so the confidence interval is always lower than 5%). This ensures statistically significant classification capability on the training set. We provide the used parameters of configuration for skewness (1.0) and for learn momentum (0.8).

B. Classifier based on SOM

Another kind of neural network self-organizing map (SOM) or Kohonen map depends on unsupervised learning. It belongs between clustering algorithm and serves as a preprocessor or a validator of detected data. The result of classification with SOM is a graph with visible clusters of same or similar samples. It can reduce multidimensional input data to a two-dimensional structure in Fig. 6.

The SOM's inner structure composes of two neurons layer. In the first layer are 10 input neurons, which use the same input vector as MLP. Each neuron in the first layer is interconnected with each neuron in the upper layer. But there are also connection between neighbor neurons in the upper layer. The weight of connections between neurons in the second layer separates different cluster groups. The learning in SOM updates neighbor connection weights. At the end of learning, network will respond to similar input samples with similar output neuron or neurons in his close neighborhood. The training of SOM (competitive learning) compute Euclidean distance for all weight vectors.

The neuron with highest excitation (i) will become the best matching unit (BMU). The BMU then adjusts towards the input vector value (p), to fit it more precisely in next training iteration as shows eq. 2.

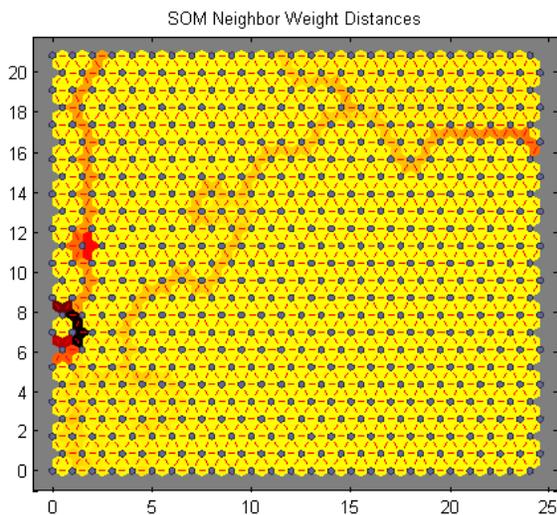


Fig. 6 SOM neighbor weight distances.

The same does neurons in close neighborhood to BMU (then eq. 2 is used for those neurons too). The parameter q is the current iteration.

$$N_i(d) = \{j, d_{ij} \leq d\}. \quad (1)$$

$$w_i(q) = w_i(q-1) + \eta(p(q) - w_i(q-1)). \quad (2)$$

The output layer has neurons in hexagonal grid 25x25. Neighborhood parameter is set to value 4, so only neurons four steps from BMU updates its weight values. The training takes 1000 iterations.

V. RESULTS

We evaluated the attack data separately for each source application (Dionaea and Tcpdump). Even if both applications run on the same hardware, there were slight differences in detected attacks. In general, both applications detected almost similar number of attacks in selected period of one month, in case of Dionaea it was 3031 attacks and for Tcpdump 3047. There were attacks in one dataset, which were not included in other dataset (i.e. not all attacks from Dionaea set is Tcpdump dataset) and vice versa. But the error rate of dataset was 0.5%, so there is no statistically significant difference between datasets.

Both application datasets show practically the same attack rate, steady in all days for whole monitoring period. They differ in delivered SIP messages. The Dionaea detects only 11286 SIP messages, Tcpdump dataset contains 228425 SIP messages. In the case of high rates of messages during attack, Dionaea is not able to detect all SIP messages correctly. There was no statistical difference in number of attacks between days of a month or days of a week. Attacks are evenly spread in time in Dionaea and Tcpdump datasets.

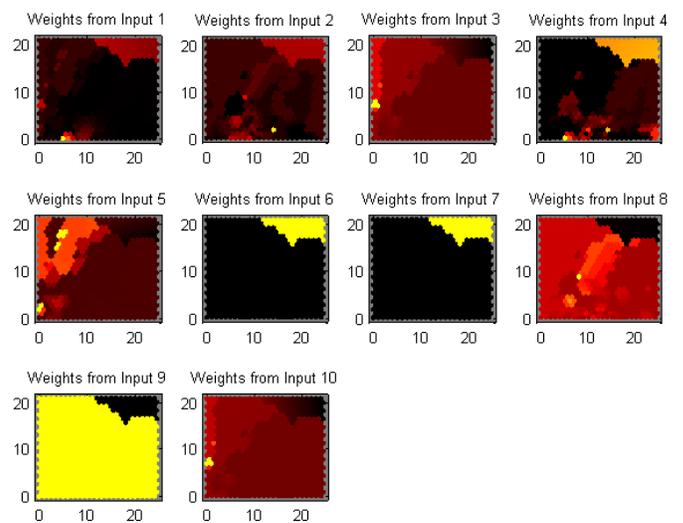


Fig. 7 Input vector weights distribution.

The error rate for Dionaea data classification was 1.287% and in Tcpdump dataset was achieved a little bit lower error

rate of 1.214%. We can claim that the SOM network is capable of finding clusters in attack data. Figure 6 shows SOM neighbor weight distances. The darker the color between neurons is, the more different the nodes are.

Figure 6 illustrates that input set contains different attack clusters. Another usefulness of SOM is in input weight evaluation of Tcpcdump dataset as is depicted in Fig. 7. It clearly shows high similarity of input parameters 6, 7, 9 and parameters 3, 10. Remaining parameters influence different part of the output layer and seem independent of each other. There is no significant difference between SOM results for Dionaea and Tcpcdump dataset.

VI. CONCLUSION

The MLP provides an accurate and reliable classifying mechanism for SIP attack classification. It enables us to detect various kind of attacks without the need to use memory or computational consuming methods. To classify attack correctly, MLP need only good representatives in the training set. The training of the MLP is time and computational consuming process. But after learning, classification is straightforward and quick, without any delays or complex computation. Each aggregated attack group is classified at the same speed, no matter which kind or how many SIP messages has an attacker used.

Single data store for attacks allows us to test new classification algorithms more quickly and with current SIP attacks. We can compare results between different algorithms or use them together to achieve higher accuracy of classification. Other methods or approach cannot adapt to a new situation easily. In some cases, like IDS based classifier, it is not possible to change whole classifier engine. In worst case, our solution needs only new parser of attack data and detection algorithm.

The SOM brings another look on input data. It clearly identifies unnecessary parameters in input vector. SOM neighborhood distances explain the problematic learning of MLP. There exist different clusters of attacks in the input set, but the differences are small. There are typical representatives of a specific attack, and other samples are alike. The variability of input parameters is low, and that complicates the training and classification. Another benefit of SOM is an input vector preprocessing. It clearly shows which parameters affects clustering. From the SOM analysis of the input set, parameters 6, 7 and 9 do not change the result of clustering. We can use different connection parameters instead of them to improve clustering.

This paper contains the description of a modular system for gathering and classifying SIP attacks. Results confirm that ANN could successfully recognize various kinds of SIP related attacks. All attacks are classified with both supervised and unsupervised algorithm. The performed SOM analysis shows limits and ways to further improvements of methods. It brings valuable information for future tuning of MLP's training set and other improvements.

REFERENCES

- [1] L. Spitzner, *Honeypots: Tracking Hackers*, Addison-Wesley Professional, 2002.
- [2] M. Collier, D. Endler, *Hacking Exposed Unified Communications & VoIP Security Secrets and Solutions*, McGraw-Hill Osborne Media, December 16, 2013, 560 p.
- [3] J. Tang, Y. Cheng, "Quick Detection of Steathy SIP Flooding Attacks in VoIP Networks," *Communications (ICC)*, 2011 IEEE International Conference on, p. 1-5.
- [4] J. Gomez, C. Gil, N. Padilla, R. Banos, C. Jimenez, "Detection of a Snort-Based Hybrid Intrusion Detection System," *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, Lecture Notes in Computer Science Volume 5518*, 2009, pp. 515-522.
- [5] M. Voznak, J. Safarik, F. Rezac, "Threat Prevention and Intrusion Detection in VoIP Infrastructures," *International Journal of Mathematics and Computers in Simulation*, Issue 1, Volume 7, 2013, pp.69-76.
- [6] F. Rezac, M. Voznak, J. Ruzicka, *Security Risk in IP Telephony*. In proceedings , CESNET Conference 2008 , 25-26.2008, Prague, ISBN 978-80-904173-0-4 (WoS)
- [7] J. Safarik, F. Rezac, M. Voznak, "Monitoring of Malicious Traffic in IP Telephony Infrastructure," *CESNET Technical Report 4/2012*, 11p., 2012.
- [8] J. Safarik, M. Voznak, F. Rezac, *Security evaluation of multimedia systems*, 28th Annual TERENA Networking Conference, TNC 2012, Reykjavik, May 2012, 10p., ISBN 978-90-77559-00-0. (SCOPUS)
- [9] M. Voznak, J. Safarik and J. Slachta, *A Neural Network Based System for Classification of Attacks in IP Telephony*, *International Journal of Circuits, Systems and Signal Processing*, Volume 8, 2014, pp. 368-375.
- [10] J. Safarik, M. Voznak, M. Mehic, P. Partila, M. Mikulec, "Neural network classifier of attacks in IP telephony," In *Proc. SPIE. 9118, Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering XII*, 91180X. Baltimore, Maryland, USA, May 22, 2014.
- [11] R. Rojas, *Neural Networks*, Springer-Verlag, 1996.
- [12] J. Heaton, *Introduction to Neural Networks for JAVA*, 2nd Edition", Heaton Research, 2008.
- [13] T. Kohonen, *Self-Organizing Maps*, Springer, 3rd edition, December 2000.

A four-state markov chain and its application in packet loss modelling for speech quality estimation of IP telephony

J. Rozhon, F. Rezac, M. Voznak, J. Safarik, J. Slachta and L. Macura

Abstract— The paper deals with modelling network effects on the quality of speech. The independent losses example presented in this paper proved the possibility of packet loss modelling based on the four-state markov chain. The main contribution of this paper lies in the implemented modelling tool and with the exploration of its possible uses as well as the measurements of the impact of the packet loss on the quality of speech and the calculation of the estimated Mean Opinion score.

Keywords—Packet Loss, Speech Quality Estimation, MOS, Jitter, Markov chains.

I. INTRODUCTION

THE growing importance of the speech and video monitoring systems, which is mainly caused by the wider use of IP-based communication, leads to increased demand for the high precision of the monitoring algorithms as well as the low computational complexity. Monitoring systems are deployed in the infrastructure of all the providers and we would like to point out that it concerns not only IP network but also mobile networks, where in fact the LTE technology brings IP communication in form of VoLTE or VoIP. The speed of algorithms assessing quality are affected by the methodology used for the quality determination. The overall result of the measurements and estimations is always a compromise of time requirements and the precision of results. In last decade many advanced mathematical approaches have appeared to improve the precision of the output results even for the quick estimation methods. One of the possible ways is described in

This research has been supported by the Ministry of Education of the Czech Republic within the project LM2010005.

J. Rozhon, F. Rezac, J. Slachta and J. Safarik are PhD. students with Dept. of Telecommunications, Technical University of Ostrava and also researchers with Dept. of Multimedia in CESNET, Zikova 4, 160 00 Prague 6, Czech Republic (e-mail: safarik@cesnet.cz, slachta@cesnet.cz, filip@cesnet.cz, rozhon@cesnet.cz).

L. Macura is a network administrator of Silesian University in Opava and he is also a researcher with Dept. of Multimedia in CESNET, Zikova 4, 160 00 Prague 6, Czech Republic (e-mail: macura@cesnet.cz).

M. Voznak is an Associate Professor with Dept. of Telecommunications, VSB-Technical University of Ostrava (17. listopadu 15, 708 33 Ostrava, Czech Rep.) and he is also a researcher with Dept. of Multimedia in CESNET (Zikova 4, 160 00 Prague 6, Czech Rep.), corresponding author provides phone: +420-603565965; e-mail: voznak@ieee.org.

this paper with a major focus on modelling network features and properties and creating the database of sets that could eventually be used for neural networks training or any similar procedure [1], [2].

II. SPEECH QUALITY ASSESSMENT

Methodologies evaluating speech quality can be subdivided into two groups according to the approach applied - conversational and listening [2]. Conversational tests are based on mutual interactive communication between two subjects through the transmission chain of the tested communication system. Listening tests do not provide such plausibility as conversational tests but they are recommended more frequently [3]. According to the method of assessment speech quality evaluation, methodologies can be subdivided into subjective methods and objective methods. To evaluate speech quality, MOS (Mean Opinion Score) scale as defined by the ITU-T recommendation P.800 is applied [4]. The basic scale of assessment as prescribed by the recommendation is depicted on Fig. 1.



Fig. 1 MOS Scale.

In order to avoid misunderstanding and incorrect interpretation of MOS values, ITU-T published recommendation P.800.1 in 2003. This recommendation defines scales both for subjective and objective methods as well as for individual conversational and listening tests.

A. Intrusive Approach

The core of intrusive (also referred to as input-to-output) measurements is the comparison of the original sample and the degraded sample affected by a transmission chain [5]. The

intrusive methods use the original voice sample as it has entered the communication system and compare it with the degraded one as it has been outputted by this transmission chain. The following list contains the most important intrusive algorithms PSQM (Perceptual Speech Quality Measurement PSQM), PAMS (Perceptual Analysis Measurement System PAMS), PESQ (Perceptual Evaluation of Speech Quality PESQ), POLQA (Perceptual Objective Listening Quality Assessment).

Among these, PESQ is currently the most commonly applied algorithm [6], [7]. It combines the advantages of PAMS (robust temporal alignment techniques) and PSQM (exact sensual perception model) and is described in ITU-T recommendation P.862. The last algorithm mentioned, P.OLQA, also known as ITU-T P.863, is intended to be a successor of the PESQ. It strives to avoid the weaknesses of the PESQ's model and to incorporate a better wideband codec analysis in comparison with PESQ. As stated above, the principle of this intrusive test is the comparison of original and degraded signals, their mathematical analysis and interpretation in the cognitive model [7].

B. Non-Intrusive Approach

Contrary to intrusive methods which require both the output (degraded) sample and the original sample, non-intrusive methods do not require the original sample. Intrusive methods are very precise but their application in real-time measurement is unsuitable because they require sending a calibrated sample and both endpoints of the examined communication. Nevertheless, we usually need to assess the speech quality in real traffic and be able to record its changes, especially degradation [7]. Two basic principles exist: a source-based approach and a priori-based. The former, the source-based approach, is based on knowledge of various types of impairments, i.e. a set of all impairments gained by comparison of original and degraded signal characteristics. The PLP (Perceptual-linear Prediction) model is a representative of this approach. PLP compares the perceptual vectors extracted from examined samples with the untainted vectors gained from original samples. As we have mentioned, it requires a database with the set of impairments and high computational complexity. Later the PLP model was modified and the computation was accelerated, nevertheless this model is not suitable for implementation in practice as its accuracy strongly depends on the quality of the database with patterns [2], [7]. As for the latter approach, I would like to mention the pioneer work of Zoran and Plakal [8]. They applied artificial neural networks (ANN) to determine statistical ties between a subjective opinion and a characteristic deformation in the received sample. They also investigated spectrograms (a spectrogram is defined as a two-dimensional graphical representation of a spectrum varying in time) and they were able to establish typical uniform aspects of speech in spectrograms. The important method was standardized in recommendation ITU-T P.562 (INMD) and in ITU-T G.107,

so-called E-model [7], [9]]. INMD measurement (In-service Non-intrusive Measurement Devices) is applied primarily to measure voice-grade parameters such as speech, noise and echo. The output from the model is a prediction of customer opinion Y_C^B (1).

$$Y_C^B = 1 + (E^B \cdot Y_{Cpre-echo}^B) \quad (1)$$

E^B is an echo and a delay multiplier, its value is between zero and one, to modify the pre-echo opinion score to take account of echo and delay impairments. $Y_{Cpre-echo}^B$ is the calculated pre-echo opinion score, on a zero-to-four scale, which takes into account effects of noise and loss. The addition of one converts Y_C^B to a one-to-five scale. All intermediate opinion score values are based on a zero-to-four scale for ease of calculation. It is possible to generate a rating R (2) using INMD measurements for a connection which is translated into a customer opinion of E-model [10], [11]. The E-model is one of the most modern method belonging to non-intrusive methods.

$$R = R_0 - I_{OLR} - I_{DD} - I_{e-eff} - I_{DTE} \quad (2)$$

R_0 is the signal-to-noise ratio at a 0 dB reference point. In the equations provided (2), the 0 dB reference point is at the 2-wire input to the telephone receiving system at the near end of the connection. I_{OLR} represents the impairment term for the overall loudness rating, I_{DD} the impairment term for the absolute one-way delay and I_{e-eff} is the impairment term for the low bit-rate coding under random packet loss conditions. The last parameter I_{DTE} represents the impairment term for the delayed talker echo. I_{OLR} represents the impairment term for the overall loudness rating, I_{DD} the impairment term for the absolute one-way delay and I_{e-eff} is the impairment term for the low bit-rate coding under random packet loss conditions. Last parameter I_{DTE} represents the impairment term for the delayed talker echo [2].

III. MODELLING THE NETWORK IMPAIRMENTS

Since the vast majority of the modern communications is performed using the technologies built upon the Internet Protocol (IP), the network impairments that can actually occur during the communication include Packet Loss, Jitter and Delay. These individual network features combine their effects on the quality of call during the transmission. And since each of them has a different nature, modelling their combined effect requires the combination of two separate models.

A. Packet Loss Modelling

The packet loss, as the name suggests, affect the call by losing one or more packets. Since each packet carries multiple samples of audio or video (for G.711 codec it is 160 samples in one packet, one sample obtains 1 Byte) the loss of those samples reflects in the quality deterioration of the call. This deterioration is as severe as much the given codec is unable to reconstruct the samples from the previous and following ones. Therefore, the impact of packet loss varies highly in

dependence on the chosen codec [1]. Throughout the development of the IP communications the various models of packet loss have been presented, starting with simple Bernoulli model, which is a two-state Markov model with a single independent probability, and ending 4-state Markov model with 5 independent transition probabilities [12]. As the models evolved, they incorporated more and more conditions of the network with the latter mentioned model incorporating the independent losses, the correlated losses, and the bursts of losses, thus being the most general model currently used [13].

Since the 4-state model is the most general model of the packet loss, the other models (Bernoulli, Gilbert, Gilbert-Elliot) are special cases of it, therefore it is the most suitable model for creating the complex packet loss modelling tool. The 4-state model is shown in Fig. 2.

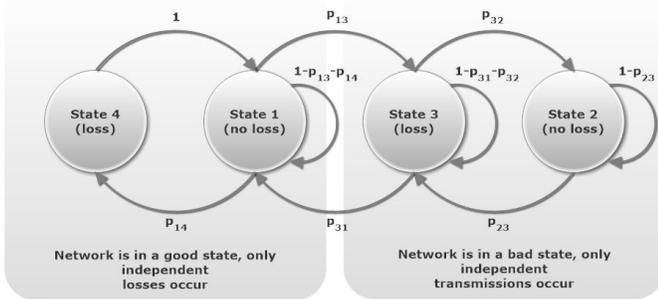


Fig. 2 The diagram of the 4-state Packet Loss Model.

The model is an actual combination of two 2-state models. The first one, which represents the good state is meant to simulate only sporadic losses, therefore, the appropriate probability should be considerably low (e.g. below 20 %). On the other hand, the losses in the bad state simulate the lengthy bursts of lost packets, where the packet loss probability should be higher (e.g. above 40 %). The system in this model passes from a good to a bad state according to the desired susceptibility of the system to failures.

Let $P(x) = \{p_1, p_2, \dots, p_n\}$ be the input sequence of packets with the length n and $P_S(x) = \{p_{S1}, p_{S2}, \dots, p_{Sn}\}$ be the sequence of ones and zeros with same length, which is a product of the aforementioned 4-state Markov model, and where 1' represents a packet loss event and 0 represents the successful transmission (no packet loss event occurred), then the output sequence of the packets can be defined as (3)

$$P_O(x) = P(x) \wedge \neg P_S(x) \quad (3)$$

The length of the output sequence is shortened by the number of lost packets.

B. Delay and Jitter Modelling

The delay and jitter are the time connected features of the packet. While delay characterizes the time needed for packet to traverse the transmission chain, the jitter is defined as the variability of the delay. The former has limited effect in today's communications, because the latencies even for long

distance calls are below the 400 ms limit defined in ITU-T G.114. Moreover, the intrusive algorithms are, in case of really constant delay, unable to compute the quality impairment, because it is not possible for them to recognize that this is a network-related issue and not the early started recording. Jitter, on the other hand, poses a great problem, because the fluent stream of packet is necessary for satisfiable communication. The jitter affects quality of call in two ways. First, countering it by de-jitter buffers increases the latencies and can contribute to exceeding the aforementioned limit. Secondly, really high jitter values (higher than the length of de-jitter buffer) lead to additional packet loss, because the late arrived packets are discarded. General consensus simulates the jitter using the normal distribution, but any other distribution can be used when appropriate. For the sake of the jitter modelling in this paper the normal distribution has been chosen with mean value equal to the desired delay and the range of jitter equal to 2.575 standard deviations of the population, which for normal distribution covers 99 % of the population.

IV. DATASET CREATION

In the previous section, the individual models of packet loss, network delay and network jitter have been introduced. These models were actually implemented using the Python language and experimental set of degraded samples has been created for the analysis. Because all the three network transmission features influence the call simultaneously, for one setting multiple rounds of modelling need to be performed with half of the attempts using the packet loss model first and timing model second, and vice versa. The main problem is the fact that neither the states and probabilities of the packet loss model nor the network delay and jitter model parameters can be measured and determined. These network transmission features can only be estimated, which means that the input data of both the models cannot be used in further analysis or the neural network inputs. Therefore, statistical substitutions have to be calculated for all the three parameters in order to be able to construct the input vectors of the neural network.

A. Packet Loss Analysis

To detect the packet loss in the stream of RTP packets, the standard procedure as defined in RFC 3550 can be used utilizing the packet sequence numbers [14]. From the received sequence of packets, the original Markov model probabilities cannot be obtained. This is because the originating state is not known, large quantity of packets can be lost in the end of the sequence and additional packet loss may have been introduced by the de-jitter buffer. However, part of the network state fingerprint is encoded in this given packet loss scheme. Since the best way to describe the loss events is using the model with the finite set of parameters, the reverse analysis of the sample in order to obtain the 4-state Markov model need to be done. This procedure is described in [12]. As the output of the Packet Loss Analysis, the parameters (transition probabilities) of the modified 4-state Markov model can be used.

B. Delay and Jitter Analysis

For the delay calculation, internal RTP timestamps can be used and the delay is therefore an easily accessible parameter, although as well as the packet loss and jitter it can only be estimated. As for the jitter calculation, the substitution of the original jitter is the interarrival jitter [14]. It is calculated in (4)

$$J(i) = J(i - 1) + \frac{|D(i - 1, i) - J(i - 1)|}{16} \quad (4)$$

where $D(i, j)$ is the difference of relative transit times for the two packets (5).

$$D(i, j) = (R_j - S_j) - (R_i - S_i) \quad (5)$$

where R_i and S_i are arrival time of the packet and the timestamp from the packet respectively.

V. HARNESSING THE OBTAINED DATA

The previous sections of this paper have provided an insight into the development of a modelling tool that has been used to create samples needed to establish a data basis for the speech quality prediction. The whole picture of the procedure used to create the data is depicted in Fig. 3.

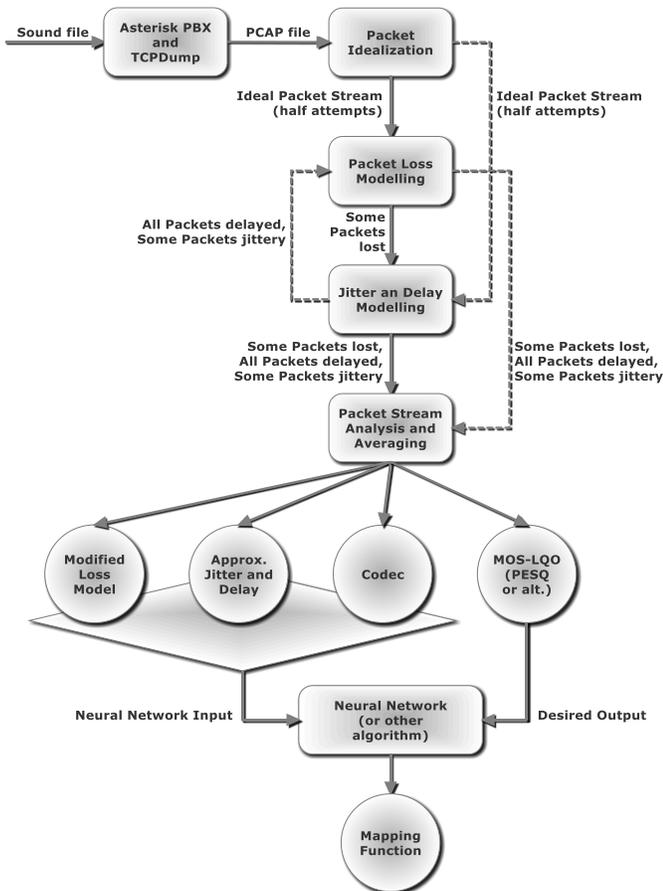


Fig. 3 Modelling Algorithm with Data Creation.

The prepared network samples (PCAP files) are first idealized to counter all the possible negative influences that

could have been introduced during the creation of the file. This process involves checking whether no packet is missing and time alignment of the packets to precise timing in conformance with the chosen codec. These idealized packets then enter the modelling tool that models the delay, jitter and packet loss as described above. The result is a modified PCAP file containing the *network fingerprint*. This PCAP file is then analyzed and the named models are then reconstructed. The reconstructed models may not resemble the original ones, but the information about the original models is lost and therefore cannot be recovered. These two new models, however, can be related to the measured quality obtained using the PESQ algorithm or any other suitable one. The relation of the network model parameters, used codec and resulting quality then can be used as an input for the mathematical mapping function that can further be used for speedup in precise voice or video quality measurements. Using the described network model, any network situation can be simulated providing the necessary data for subsequent analysis, which ties together the fingerprint of the network, used codec and resulting speech or video quality for further use in any suitable mathematical mapping function. Experimental Measurements. As a part of the conducted research several experimental measurements have been performed to solidify the assumptions stated above. In this section, the possibilities of using the neural network to estimate the quality of speech based on the packet loss data will be discussed. For the experiment, two minimalistic network designs have been created. The first one is the model of the wired network with independent losses only. The second one then models the wired network with dependent losses. Both of the models use G.711 (A-law) encoded speech data, but for the sake of this paper the former one will be discussed in greater detail to present the proposed approach.

A. Modelling and Estimating the Independent Losses

For the model presented in the Fig. 2 to behave as a simple Bernoulli model with just one independent variable, it is necessary to set the transition probabilities as follows (6):

$$\begin{aligned} p_{41} &= p_{23} = p_{31} = 1, \\ p_{13} &= p_{32} = 0 \end{aligned} \quad (6)$$

These constraints come from the fact that only two of four model states are used and that the starting state can be any given one. Therefore, setting the probabilities as it is shown ensures that after many transitions the model will converge into the Bernoulli one. The only transition probability not listed p_{14} is the independent variable of this model. Since the model needs to be in a steady state to generate the expected outcome, the first 1000 transitions are discarded. As long as the model works with the precision in percents, this value is sufficient. On the listening side (output of the model) however, the information about the model's transition probabilities cannot be obtained. Therefore, the generally used term - packet loss probability P_{pl} is observed. For this experiment the range for the P_{pl} is set from 0 to 0.15 with the step of 0.01, which ensures the results in whole spectrum of the MOS range.

From the balance equations for the given model the p_{14} can then be calculated as follows (7):

$$p_{14} = \frac{\pi_4}{1 - \pi_4} \wedge \pi_4 = P_{pl} \tag{7}$$

With the model set, input data have been modified based on the models output resulting in degraded voice samples, which have then been used for the MOS_{LQO} calculations. The results of these calculations are in the chart in the Fig. 4.

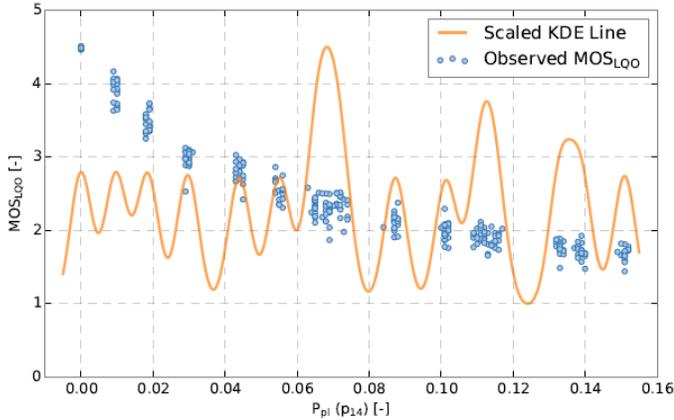


Fig. 4 MOS_{LQO} vs. PL Probability and respective KDE Line.

These observations display an obvious exponential decrease in speech quality. However, due to the probabilistic nature of the model, the observations are shifted from the input coordinate P_{pl} has been chosen with a 0.01 step, therefore, the data cannot be easily clustered. To cluster the data for the further analysis, histograms or kernel density estimation can be used, since the distribution of the observations on the x-axis is the set of 1D data. Due to the obvious feature of the observations, which form wider and tighter clusters, the histogram approach could lead to wrong classification. For this reason, kernel density estimation have been used with gaussian-shaped kernels. The resulting curve is shown in the scaled and shifted manner in the Fig. 4. By exploring this curve, the clusters of observations can easily be identified when looking on the maxima of the curve as the clusters' centers of gravity and minima as the clusters' boundaries. Since the clusters of observations have been identified, the descriptive parameters of these clusters can be calculated, such as the outliers, means and standard deviations. Based on these parameters the neural network can be trained. However, to achieve higher precision one more step of data preprocessing have been added and the fitting curves for the mean (8) and standard deviations have been calculated.

$$MOS_{LQO} = 2.805 \cdot e^{-20.989 \cdot P_{pl}} + 1.622 \tag{8}$$

The outliers, means, standard deviations and fitting curves are shown in the Fig. 5. Although the fitting curves provide sufficient way to estimate the MOS_{LQO} based on the percentage of the independently lost packets, the estimation system is meant to be much more robust. Therefore, the estimation is done using the neural networks. In this simple case, where the

exponential function is to be estimated, the network with only two hidden neurons can perform well.

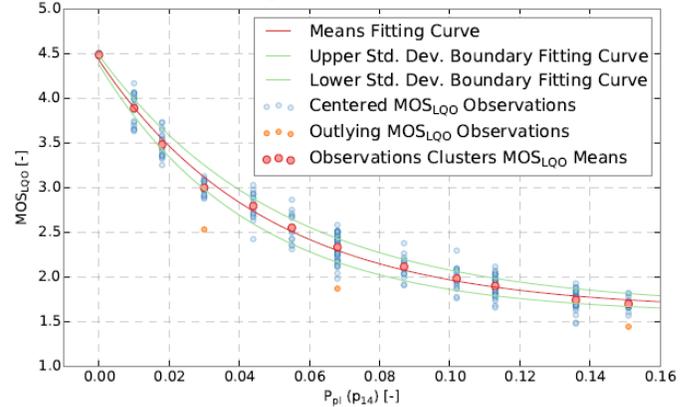


Fig. 5 MOS_{LQO} Observations Clusters and Fitting Curves.

This is because of the similarities between exponential curve and the sigmoid function used in the neurons. The comparison between the fitting curve of the means of the observations and the neural network estimate of it is shown in the Fig. 6.

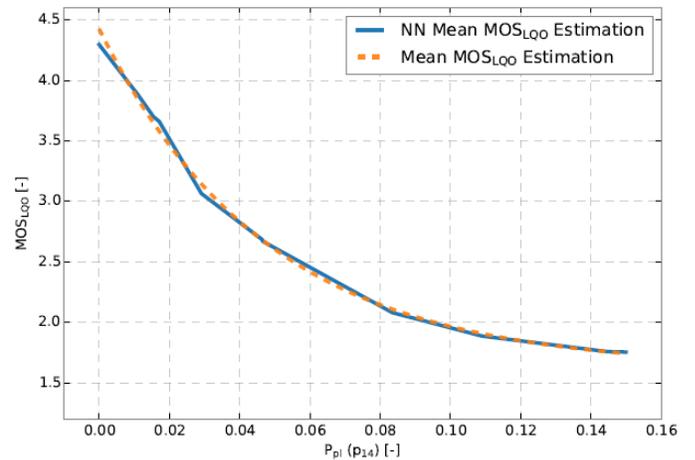


Fig. 6 The Comparison of the NN based MOS_{LQO} Estimation with Fitting Curve of Means.

The comparison of the two curves shows the neural network approximates the fitting curve almost perfectly and can therefore be used for the MOS_{LQO} calculations in the network with independent losses.

B. Modelling Dependent Losses

Unlike the independent losses the dependent ones are much more complex problem. When talking in the terms of the presented 4-state model, to model the dependent losses only the states 1 and 3 are to be used. Similarly to previous subsection, the constraints defining the model are as follows.

$$\begin{aligned} p_{41} = p_{23} &= 1, \\ p_{14} = p_{32} &= 0 \end{aligned} \tag{9}$$

The remaining transition probabilities p_{13} and p_{31} are the independent variables of the model. The observer, as well as

for the previous case, cannot determine these probabilities based on the output packet stream, therefore the analytically obtained variables have to be used. These include P_{pl} - packet loss probability, P_{bpl} - packet loss probability in burst, ρ - density of packet loss in bursts. These variables are conform to ones defined in [12]. In the experiment the total percentage of lost packets again have ranged from 0 to 0.15 with a step of 0.01, but since there are two independent variables the input sets were chosen (from the whole interval 0-1) so that both probabilities are in whole percents. The equation specifying the relation between probability of packet loss P_{pl} and the variables p_{13} and p_{31} can again be derived from the balance equations of the model and looks as follows (10).

$$\pi_3 = P_{pl} = \frac{p_{13}}{p_{13} + p_{31}} \quad (10)$$

Based on this input data, the wired network with dependent losses has been modeled resulting in the set of observations as shown in Fig. 7. In this chart the MOS_{LQO} values are encoded in color of the scatter points. It can be observed that with increasing burstiness of the losses the voice quality increases, thus showing the evident relation between speech quality and burstiness of the packet losses.

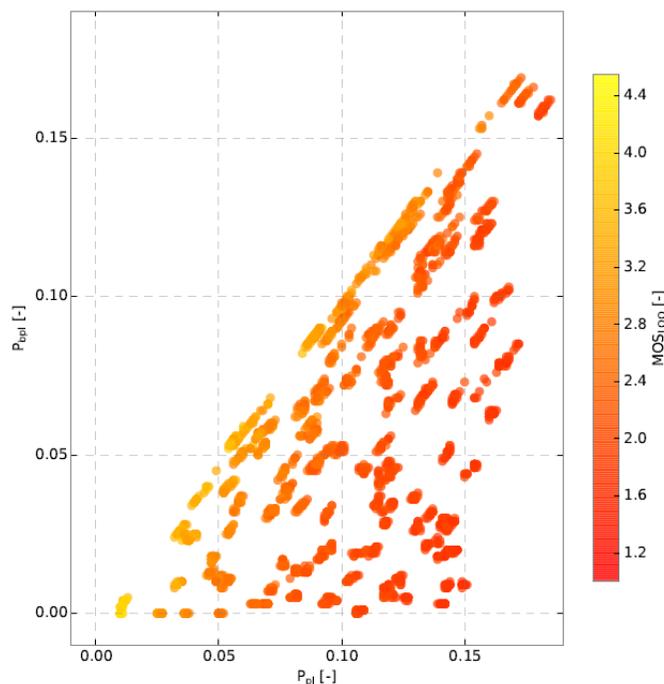


Fig. 8 The color-encoded MOS_{LQO} values in relation to overall and burst Packet Loss.

This relation can again be encoded in the structure of the neural network. However, there are several obstacles on the way and therefore this is the main aim of our current research.

VI. CONCLUSION

In this paper the approach to modelling network effects on the quality of speech and possibly of video as well has been

presented. The modelling tool harnessing this approach have been implemented and several measurements proving the viability of the approach have been conducted. The independent losses example presented in this paper proved the possibility of estimating the speech quality based on the network packet loss. The main contribution of this paper lies with the implemented modelling tool and with the exploration of its possible uses as well as the measurements of the impact of the packet loss on the quality of speech. The detailed process of evaluating the measurement data and creating the training sets for the neural networks described in this paper is a great asset for the network administrators and telco providers and shifts the topic of speech quality estimation further. In the future work the research will be focused on studying the consecutive losses and combination of both models, which will form the solid basis for the robust quality estimation tool.

REFERENCES

- [1] Q. Fu, K. Yi, M. Sun, "Speech quality objective assessment using neural network," IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000, pp. 1511-1514.
- [2] M. Voznak, "Recent advances in speech quality assessment and their implementation," Lecture Notes in Electrical Engineering, 282 LNEE, 2014, pp. 1-14.
- [3] A. E. Mahdi, D. Picovici, "Advances in voice quality measurement in modern telecommunications," Digital Signal Processing, Volume 19, Issue 1, January 2009, pp.79-103.
- [4] Methods for subjective determination of transmission quality, ITU-T Recommendation P.800, Geneva, 08/1996.
- [5] A. Rix, M. Hollier, A. Hekstra, J. Beerends, "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment," AES: Journal of the Audio Engineering Society, Volume 50, Issue 10, 2002, pp. 755-764.
- [6] Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, ITU-T Recommendation P.862, Geneva, 02/2001.
- [7] M. Voznak, Non-intrusive Speech Quality Assessment in Simplified E-Model, WSEAS Transactions on Systems, Issue 8, Volume 11, 2012, pp.315-325.
- [8] J. Palakal, M. Zoran, "Feature extraction form speech spectrograms using multi-layered network models," In Proc. IEEE International Workshop on Tools for Artificial Intelligence, Architectures, Languages and Algorithms, 1989, pp. 224-230.
- [9] The E-model: A computational model for use in transmission planning, ITU-T Recommendation G.107, Geneva, 04/2009.
- [10] M. Voznak, E-model modification for case of cascade codecs arrangement, International Journal of Mathematical Models and Methods in Applied Sciences, Volume 5, Issue 8, 2011, pp. 1439-1447.
- [11] H. Assem, D. Malone, J. Dunne, P. O'Sullivan, "Monitoring VoIP Call Quality Using Improved Simplified E-model," 2013 International Conference on Computing, Networking and Communications, ICNC 2013, art. no. 6504214, pp. 927-931.
- [12] S. Salsano, F. Ludovici, A. Ordine, A. Giannuzzi, "Definition of a general and intuitive loss model for packet networks and its implementation in the Netem module in the Linux kernel," University of Rome Tor Vergata, Technical Report, 70 p., 2012.
- [13] A. Jurgelionis, J. Laulajainen, M. Hirvonen, A. Wang, "An empirical study of NetEm network emulation functionalities," Proceedings - International Conference on Computer Communications and Networks, ICCCN, art. no. 6005933, 2011.
- [14] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, RFC3550 - RTP: A Transport Protocol for Real-Time Applications. IETF, July 2003.

Experimental analysis of the effects of turbulent jets in shallow water bodies

Robles L. Isidro, Palacio P. Arturo, Rodríguez V. Alejandro

Abstract— in this work several series of turbulent experiments produced by the injection of water in the form of shallow jets into the tank filled with water of the same density were conducted. Dye was injected with the source fluid as tracer. The concentration of the dye in the shallow turbulent flow was determined using a video imaging technique. The present laboratory experiments were conducted in a tank of small thickness, however, was significantly widened to avoid the effect of the side walls. The space between the parallel walls of the tank can be varied during the experiments. The large-scale turbulent flow in the small space between the walls of the tank is confined to essentially two-dimensional motion, and the motion is retarded by the force of friction. The effect of the friction, evaluated was found to have an important effect on the entrainment processes. These findings are useful for turbulent modeling of the shallow shear flow and on its application to the large scale heat and mass exchange processes in the lagoons, lakes, oceans and atmosphere.

Keywords—Jets, shallow flows, turbulence, video imaging technique.

I. INTRODUCTION

SOME important heat and mass exchange processes in rivers, lakes and oceans are associated with large-scale quasi-two-dimensional turbulent motion of shallow depth. Possibly the most important example of all shallow shear flows is the Gulf Stream in the Atlantic Ocean. With a width L of 60 km and depth h of 600 m, the horizontal to vertical length scale ratio of the Gulf Stream is also 100 to 1. The horizontal turbulent viscosity ν_T in the stream is estimated to be less than $100 \text{ m}^2/\text{s}$ [1]. The maximum speed in the stream, V , is 2 m/s. Thus the dimensionless eddy-viscosity coefficient of the Gulf stream, $\nu_T/VL \sim 0.001$, is two orders of magnitude smaller than the typical value $\nu_T/VL \sim 0.1$ found in free shear flows, such as in jets and mixing layers.

The quasi-two-dimensional turbulent flow consisting of large scale and the small scale turbulent motion has been the subject of a number of recent investigations. In this case, the main motion of large scale is confined to move in a predominantly horizontal direction between the free surface and the channel bed. The small movement, on the other hand, with a length scale less than the depth of the flow is three-dimensional, since it is free to move in all directions.

Although attempts have been made to compare model simulations of shallow turbulent flows with experimental data [2], very few laboratory experiments of shallow turbulent flows were conducted in small depth. Recently, we were able

to produce turbulent flows in a tank of small depth. The tank, as shown in Fig. 1, consisted of a pair of parallel walls (235 cm wide and 110 cm high) with a small space between the walls. The dye concentration in the flow was measured by a video imaging technique.

This kind of studies allows obtaining information useful for turbulent modeling of the shallow shear flow.

II. EXPERIMENTAL SET

In the experimental set we used a tank with a large lateral length scale compared with the depth. It had a ratio h/L of 4.3×10^{-5} approximately. Quasi-two-dimensional flow produced in the tank depends on the momentum flux and buoyancy flux from the source. In the first series of the jet experiments, we take care to ensure that the water temperature of the source and the tank were equal within a measurable limit of 0.05 degree centigrade (Fig. 2). So, the jet produced in the tank therefore is not affected by buoyancy force.

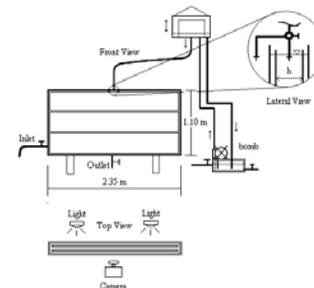


Fig. 1 Experimental set.

The momentum flux is nearly a constant in the test TW1 (wide test). In this case, the depth of the flow $h = 4.4 \text{ cm}$ is large and the wall friction is negligible. This kind of turbulent flow is as show in figures 10a-d.

The initial development characterized by the formation of a head which is significantly greater than the jet behind. Upon impingement of the head on the bottom of the tank, the jet splits up and glides along the floor. The turbulent flow then moves up along the left and right side walls. The flows of the jets in tests TN1, TN2, TM1, TM2 and TM3 are affected by friction. Friction is most important in tests TN1 and TN2 (narrow tests). The head of the jets in these cases stays in the mid-depth and is not able to penetrate deep enough to hit the floor (Fig. 6). The penetration is about equal to one friction

length scale which is estimated in these tests to be $h/cf \sim 75$ cm if $cf = 0.008$. The depth of the water in the tank is 110 cm.

In order to minimize the deflection of the walls, the tank was constructed of a double wall structure. The walls, of the inner tank were kept perfectly parallel to each other by filling both the inner and outer tank with water. Wall deflection was eliminated since the net hydrostatic pressure on the inner tank walls is zero.

The outer tank walls are 125 cm high and 245 cm wide. The inner tank walls are 110 cm and 235 cm, respectively. The distance between the parallel walls in the inner tank are 4.4 cm (W), 1.29 cm (M), and 0.60 cm (N), series of tests. Table I summarizes the condition of the experiments.

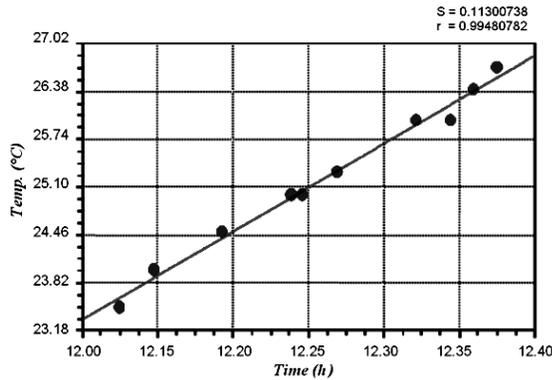


Fig. 2 Increase temperature of injection water due the bomb.

Blue dye of known concentration was used as a tracer. The turbulent flow was illuminated by back light and recorded by a video camera (SONY 3CCD) at a rate of 30 frames per second. The video images were subsequently digitized frame by frame and analyzed using a computer.

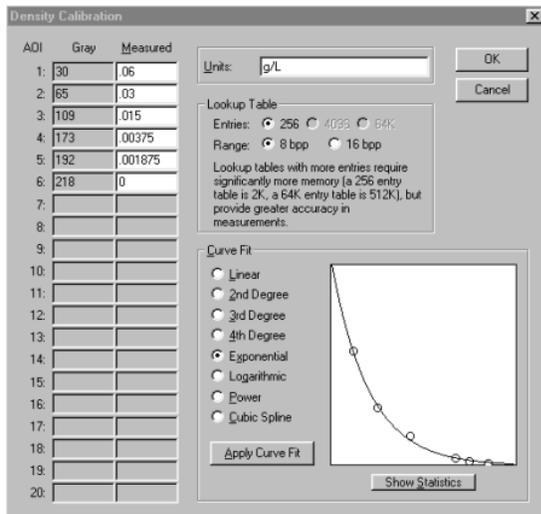


Fig. 3 Calibration curve.

Table I Summary of test conditions.

Test	h (cm)	Qo (cm ³ /s)	Co (g/l)
TN1	0.6	76.48	0.125
TN2	0.6	62.31	0.125
TM1	1.29	101.94	0.6
TM2	1.29	101.94	0.06
TM3	1.29	92.30	0.06
TW1	4.4	68.98	0.06

The video images are stored as 24 BPP in BMP format. There are 640 pixels x 480 pixels in a frame and each pixel has three basic colors, red, green, and blue, each with values varying from 0 to 255 (table II). The RGB values in the BMP file are proportional to the intensities of red, green, and blue light through the video camera and are related to the concentration of the dye in the turbulent flow through a calibration curve (Fig. 3).

The calibration of the video camera was conducted using diluted samples of known dye concentration. Video images were taken of the samples in a small Plexiglas box of the same thickness as the inner tank and under the similar lighting condition as the experiment (Fig. 4).

Table II Color coding

red	green	Blue	result
0	0	0	Black
255	255	255	White
255	0	0	Pure red
0	255	0	Pure green
0	0	255	Pure blue
X	X	X	Any color

The non-linear exponential relationship is used to relate the relative dye-concentration, c/c_0 with the p -value.

The concentration of the dye in the turbulent flows of the jet is determined for each of the 640 pixels x 480 pixels in the image file. Since the light intensity over the 110 cm x 235 cm area of the tank is not exactly uniform, the dye concentration is determined through the change in light intensity relative to the light in the background (table 3).

The light intensity of the background was not exactly uniform. However, by using the p -value to measure the relative light intensity, the calibration in the central region of the tank was found to be not significantly different from the calibration obtained from elsewhere in the tank.

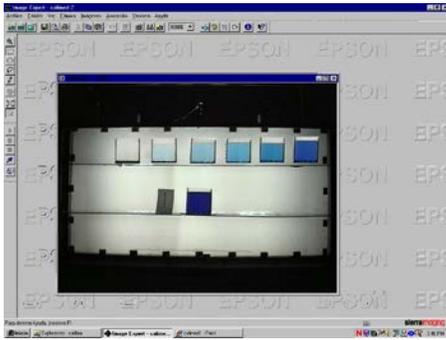


Fig. 4 Diluted samples of know dye concentration.

Therefore, the dye concentration in the turbulent flow presented here was calculated based on the calibration curve obtained from a small area located in the central region of the tank.

Table III Results of test TN1.

No.	Dye (g)	Vol (L)	Acu. Vol.(L)	Co (g/L)	Color coding
1	0.125	1	1	0.1250	33
2	0.125	2	3	0.0417	107
3	0.125	2	5	0.0250	133
4	0.125	2	7	0.0179	149
5	0.125	2	9	0.0139	157
6	0.125	2	11	0.0114	164
7	0.125	2	13	0.0096	174
8	0.125	2	15	0.0083	177
9	0.125	2	17	0.0074	173
10	0.125	2	19	0.0066	176
11	0.125	2	21	0.0060	189
12	0.125	2	23	0.0054	193
13	0.125	2	25	0.0050	184
14	0.125	2	27	0.0046	204

III. CONFINEMENT EFFECT

Without consider the dynamical dependence of the flow on buoyancy and friction forces, turbulent motion is expected to be affected by the kinematic constraint of the walls. The energy cascade process in quasi-two-dimensional turbulent flow of shallow depth is expected to be different from the process in unconfined three-dimensional turbulent motion. Vortex stretching, the dominant mechanism in three dimensional turbulence, is absent from the process in two-dimensional motion. The kinematics of the confinement effect on the turbulent flow of shallow depth has been the subject of a number of recent investigations, [3], [4], and [5]. The results, however, were not conclusive.

IV. FRICTION EFFECT

Whereas the confinement effect may be negligible, the friction effect is not.

The friction effect suppresses large-scale motion. The results are smallest entrainment rate and higher dye concentration. The limiting case of the jet is of considerable significance. In the present context, jet free refers to the case when friction effect is negligible. We may speak of free jets of two kinds. The free jet of the first kind (often referred to as the plane jet) is unconfined and free of friction effect but, to the best our knowledge, there have not been any experiments conducted for the starting free yet of the first kind. The jets in test TW1 are free jets of the second kind; the jets in this test are confined but the friction effect is again negligible.

V. INSTANTANEOUS CONCENTRATION PROFILES

The advantage of the video imaging method is the ability of the method to find the concentration profile of the entire flow field at any instant using one frame from a sequence only the data one point at a time. A series of instantaneous concentration profiles was obtained at a cross section $x = 30$ cm across the jet of test TW1, are show in the Fig. 5. In the case of the tests TN1 and TM1 which have shallower depths. As the depth of the turbulent flow becomes shallower, the turbulent motion observed from these instantaneous profiles becomes progressively more orderly. The orderly structure of the turbulent motion is closely associated with the reduction of entrainment and mixing in the flow.

The test TN1 and TN2 are strongly affected by the friction effect. Fig. 6 shows how the head of the starting jet in test TN1 is arrested by friction at a location approximately equal to one friction length scale (that is, at $x \approx h/cf$). As the advancing speed of the front is diminishing, the turbulent fluid in the head is pushed sideways in a lateral direction.

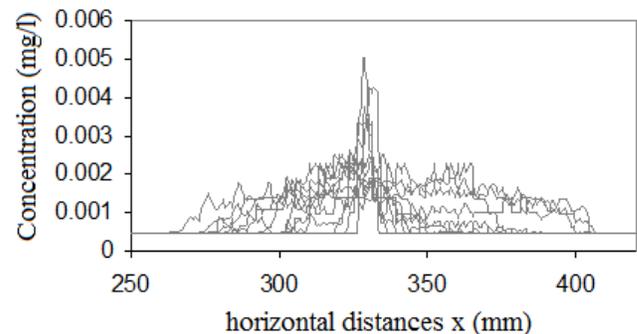


Fig. 5 Instantaneous concentration profiles of test TW1.

Eddies from the jet behind the head are continuous being fed into the head. Much effort is apparently required to force-feed the eddy into the head. This is evident by the stepwise increase in the width of the head and the meandering motion of the turbulent jet as show in Fig. 6.

Both entrainment and meandering of the turbulent jet contributes to the lateral spreading of the jet. The entrainment process is a direct contribution to the dilution of the tracer concentration. The meandering process although increases the

width of mean concentration profile, does not have a direct effect on jet dilution.

VI. STARTING JETS

The Initial development of turbulent flow produced by the momentum source is examined in this chapter as a starting jet. In general, the starting jet is characterized by the formation of a ‘head’ which is quite large in size compared with the ‘jet’ of turbulent fluid that follows the head, as shown as a series of images in Fig. 7.



Fig. 6 The head of the starting jet in test TN1 is arrested by friction.

The views of the starting jets are shown as a series of images in figures 8a-8d, 9a-9d and 10a-10d, for the tests TN1, TM1 and TW1 respectively. The confinement and friction effects on these starting jets depend on the depths of the flow which are $h = 0.60$ cm, $h = 1.29$ cm, and $h = 4.4$ cm, for the three groups of tests (a) TN1, TN2, (b) TM1, TN2, TM3, and (c) TW1, respectively. The friction effect is likely to be negligible in the test TW with the greatest depth of $h = 4.4$ cm, but is expected to be significant in the test TN1, TN2, with the smallest depth of $h = 0.60$ cm.

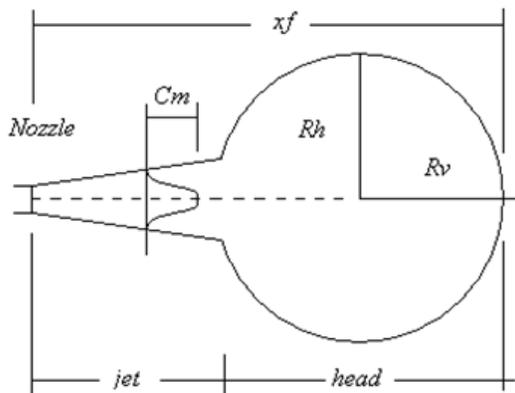


Fig. 7 the head and jet regions of the starting jet.

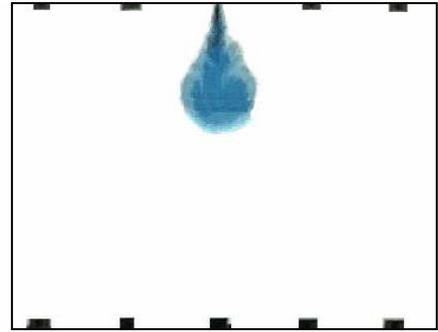


Fig. 8a Test TN1: $h=0.60$ cm, $t= 3$ s

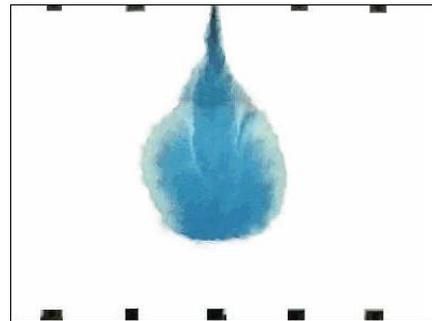


Fig. 8b Test TN1: $h=0.60$ cm, $t= 8$ s

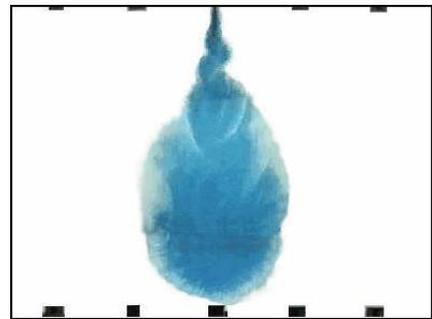


Fig. 8c Test TN1: $h=0.60$ cm, $t= 18$ s

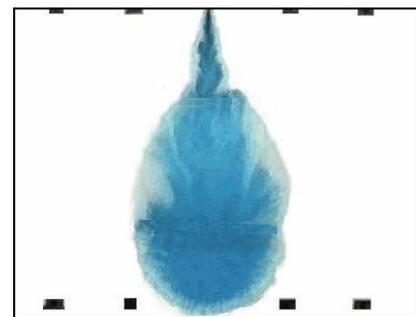


Fig. 8d Test TN1: $h=0.60$ cm, $t= 35$ s

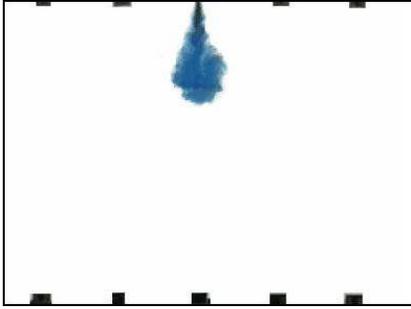


Fig. 9a Test TM1: $h=1.29$ cm, $t= 3$ s

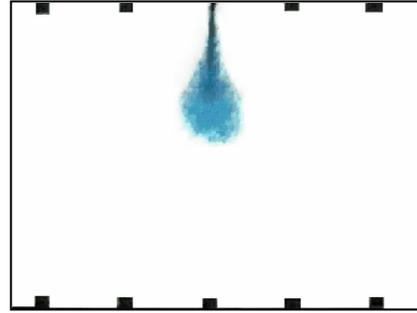


Fig. 10a Test TW1: $h=4.4$ cm, $t= 3$ s

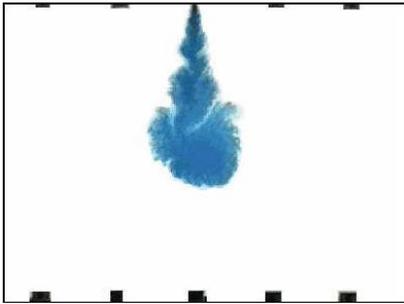


Fig. 9b Test TM1: $h=1.19$ cm, $t= 8$ s

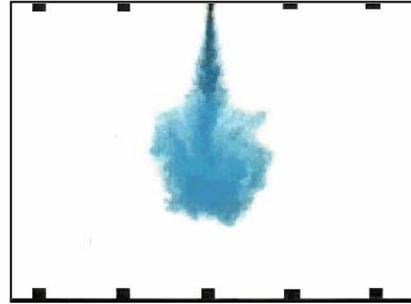


Fig. 10b Test TW1: $h=4.4$ cm, $t= 8$ s

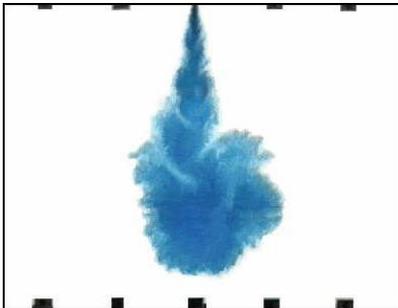


Fig. 9c Test TM1: $h=1.29$ cm, $t= 18$ s

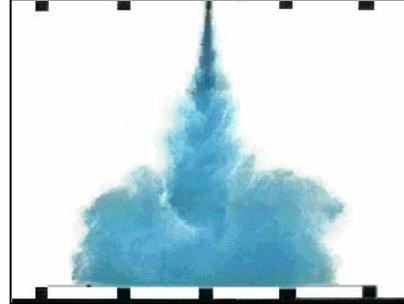


Fig. 10c Test TW1: $h=4.4$ cm, $t= 18$ s

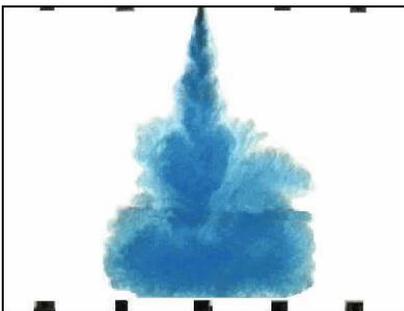


Fig. 9d Test TM1: $h=1.29$ cm, $t= 35$ s

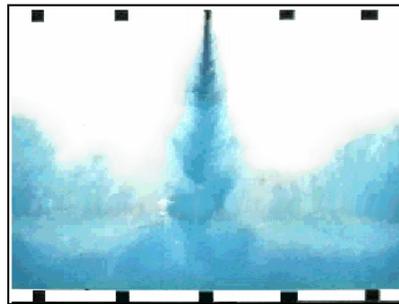


Fig. 10d Test TW1: $h=4.4$ cm, $t= 35$ s

VII. NEAR-FIELD CONFINEMENT EFFECT

The initial development of the turbulent jet is relatively independent of the friction effect (1). The inverse-square concentration data follow the linear relations

$$\frac{C_s^2}{C_m^2} = 0.64 \ xf \ \frac{Cf}{h} \tag{1}$$

Where C_s is concentrations scale and C_m is concentration maximum, Cf is the friction coefficient, h distance between walls, and xf is the distance from the head of turbulent flow.

VIII. FAR-FIELD FRICTION EFFECT

Mixing diminishes in the far field region so that the dye concentration approaches approximately the asymptotic values (2). The experimental data are scattered because they are the instantaneous values of an unsteady turbulent flow.

$$\frac{C_s^2}{C_m^2} = 0.028 \tag{2}$$

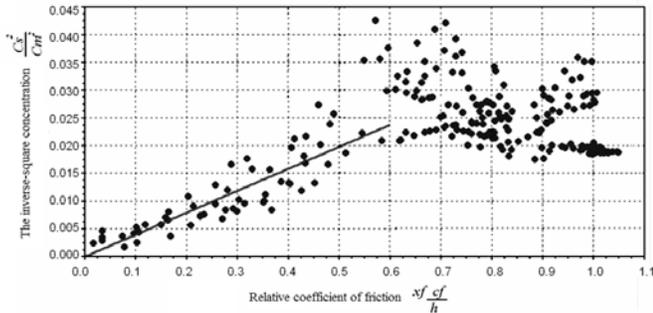


Fig. 11 Results of dye concentration maximum C_m across widest section of head. Data are inverse square of C_m normalized by concentrations scale C_s .

IX. CONCLUSIONS

The results were conducted in a tank of variable thickness. The tracer concentration distributions in these turbulent flows were determined by using a video imaging technique. The friction effect was significant in the first series (N-series and M-series) of the experiments. The confinement of the flows in these experiments to the depths of 0.60 cm and 1.29 cm led to a body-force effect allowing the energy of large-scale turbulence to be removed at a rate directly proportional to friction. In the last series (the W-series) of the experiments, the tank thickness was 4.4 cm. Friction effect was negligible in this case, but the flow was confined because the lateral length scale of turbulent motion was quite large compare with the depth of the flow. Despite the confinement of the large-scale

turbulence in the small space of 4.4 cm between the walls, the turbulent flow was observed to behave as a free turbulent flow of much greater thickness. The main result is that only the turbulent flows in the far field ($xf \ Cf/h > 0.5-0.6$) are dependent on the friction effect (Fig. 11)..

REFERENCES

- [1] Stommel, H. (1958) The Gulf Stream. University of California Press, Berkeley, 202 pp.
- [2] Babarutsi, S., Nassiri, M., and Chu, V. H. (1996) "Computation of shallow recirculating flow dominated by friction" J. Hydr. Engrg., ASCE, 122(7), 367-372.
- [3] Chu, V.H. and Babarutsi, S. (1988) "Confinement and bed-friction effects in shallow turbulent mixing layers," J. of Hydraulics Engineering, ASCE, Vol. 114, pp. 1257-1274.
- [4] Chu, V.H. and Baines, W.D. (1989) "Entrainment by a buoyant jet between confined walls" J. of Hydraulics Engineering, ASCE, Vol. 115, No. 4, pp. 475-492.
- [5] Babarutsi, S., and Chu, V.H. (1991) "A two length-scale model for quasi-two-dimensional turbulent shear flows," Proc. of the 24th IAHR Congress, Madrid, Vol. C, pp. 51-60.

Design of M2M Service Capability for Access to Location Information

Ivaylo I. Atanasov and Evelina N. Pencheva

Abstract—Service capabilities in Machine-to-Machine (M2M) communications provide data mediation functions that may be shared by different applications through application programming interfaces. The paper presents an approach to design RESTful Web Services for access to location information of M2M devices. M2M location information is modeled as a tree resource structure where resources are manipulated by REpresentational State Transfer (REST) primitives. Web Services performance characteristics are evaluated by simulation.

Keywords—Machine-to-Machine communications, Mobility Service Capability, Web Services, Service Level Agreements.

I. INTRODUCTION

MACHINE-TO-MACHINE (M2M) communications allow intelligent objects, which are uniquely identified, to capture data from or to control the environment, and to exchange information without human intervention. It is expected that M2M communications will change the businesses like intelligent transport systems, healthcare applications, city automation, energy efficiency, etc. Forecasts estimate that by the 2018 M2M connected devices will be 19.7% of all connected devices and the mobile M2M communications will represent 6% of all mobile data [1]. This means that the network traffic patterns will be changed dramatically.

One of the main challenges standing in front of M2M technologies is the fragmentation of solutions. Most of the solutions developed and implemented to date address specific application requirements which results in heterogeneous forms of technologies, platforms, and data models [2], [3], [4]. There is a lack of interoperability. It is necessary to outline capabilities that can be shared between different applications. Service capabilities are software modules that are exposed to M2M applications through the use of application programming interfaces (APIs). The Web Services (WS) model is based on the assumption that communications are like invocation of remote service whose nature can be ignored. But this model is not suitable for M2M applications, as M2M devices possess constrained resources with tangible states that can be

manipulated. REST (REpresentational State Transfer) is adopted as a method for M2M modeling and implementation. In REST, each physical or logical entity is represented as a resource which has particular state. The resource can be addressed through HTTP Uniform Resource Identifier (URI) and its states can be retrieved and updated. Resources can be created and destroyed respectively.

The related research concerning access to location information addresses specific solutions, but does not study the necessary generic functionality [5], [6]. The related implementations deal with positioning methods and usage of location data, but do not concern programmability issues.

The traffic models and overload control mechanisms developed for human driven communications are not applicable to M2M communications [7], [8]. The last ones feature small and infrequent data transmissions in contrast to web browsing, file transfer and variable bit rate streams. Most researchers in the area focus on overload mechanisms which are radio technology specific [9], [10], [11], [12], [13]. While it is important to dimension the network that provides connectivity for M2M devices, there is also a need to evaluate the traffic generated by M2M applications at application level in order to define Service Level Agreements (SLA) for different M2M service providers. Service Capability Server (SCS) may provide one or more M2M service capabilities [14], [15]. SCS appears to be a possible bottleneck when deploying M2M applications.

In this paper, we propose an approach to design RESTful WS for access to location information and evaluate some performance characteristics concerning WS deployment.

The paper is organized as follows. Next section describes typical use case scenarios, which are used to identify the required WS functionality for access to location information of M2M devices. Section III presents resources modeling location data. Resources are organized in a tree structure and can be uniquely identified. Section IV describes a model of SCS which provides access to M2M location data. The SCS model applies access control based on SLA in order to prevent congestion. Simulation parameters and results are discussed in Section V. The conclusion summarizes the contribution.

II. USE CASES FOR ACCESS TO M2M LOCATION DATA

The accuracy of location information depends on the application. Additionally, location information may be reported periodically or on demand, and notifications of

I. I. Atanasov is with the Faculty of Telecommunications, Technical University of Sofia, Bulgaria (phone: +359 2 965 2050; e-mail: iia@tu-sofia.bg).

E. N. Pencheva is with the Faculty of Telecommunications, Technical University of Sofia, Bulgaria (e-mail: enp@tu-sofia.bg).

distance changes between monitored devices may also be available to applications. Periodic location reporting as well as triggered location reporting are provided in case of available subscriptions which require subscription management functionality.

M2M applications run service logic and use available service capabilities. M2M applications may reside in the network (NA), in an M2M gateway (GA) or in M2M device (DA).

Fig.1 illustrates the message flow when an NA makes a query for M2M device location. Fig.2 illustrates the message flow for subscription to notifications triggered when an M2M device comes in (or goes out of) a specific area, triggered notifications and subscription termination. Fig.3 illustrates the message flow for subscription to period notifications about M2M device location, periodic reporting and subscription ending.

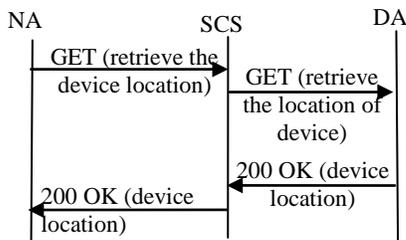


Fig.1. Query for the device location

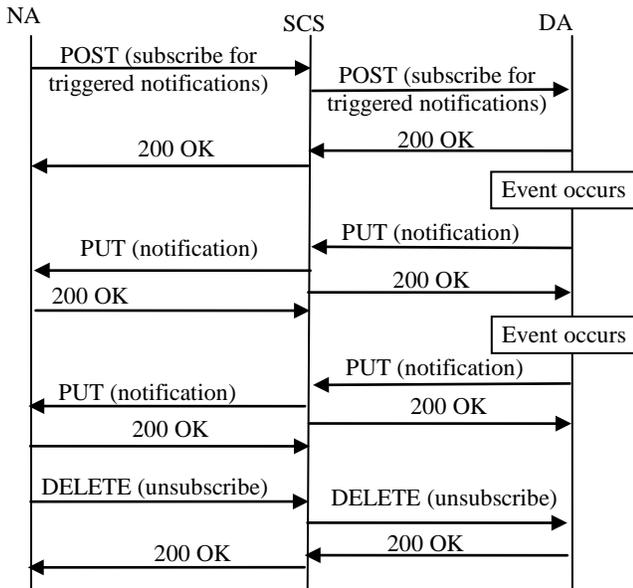


Fig.2. Triggered device location notification

An NA may retrieve the location for an M2M device using the GetDeviceLocation operation. In order to subscribe for triggered location change notifications, the NA uses StartTriggeredLocationNotification. An DA uses TriggeredLocationNotification to notify the NA about a change in the location of the M2M device. Using the StopTriggeredLocationNotification operation the NA may stop notifications. Similarly, for periodic location reporting StartPeriodicLocationNotification, TriggeredLocationNotification and

StopTriggeredLocationNotification operations are used. Some applications may be interested in device distance from a location. Some network applications may be interested in device distance from a location and the GetDeviceDistance operation may be used for this purpose.

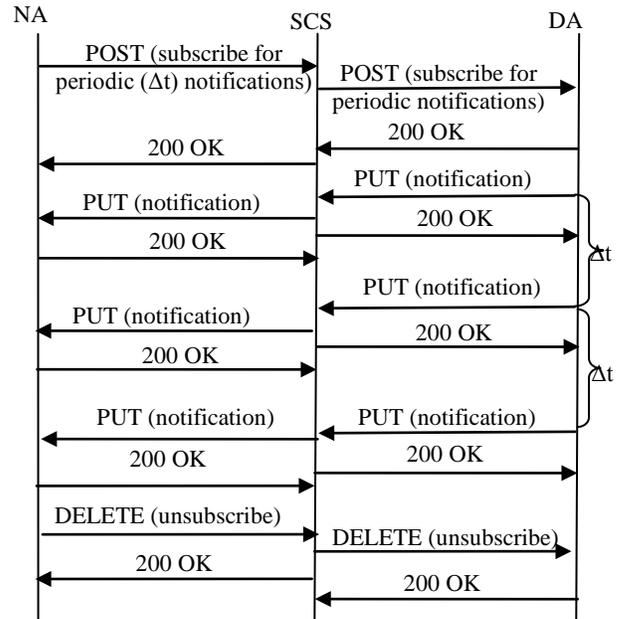


Fig.3. Periodic device location notification

Table 1 lists some of the RESTful WS operations used for access to location information of M2M devices and the respective HTTP methods that can be used for their implementation.

TABLE I. LOCATION WEB SERVICE OPERATIONS AND THE CORRESPONDING HTTP METHODS

Web service operation	HTTP method
GetDeviceLocation	GET
StartTriggeredLocationNotification	POST
TriggeredLocationNotification	PUT
StopTriggeredLocationNotification	DELETE
StartPeriodicLocationNotification	POST
PeriodicLocationNotification	PUT
StopPeriodicLocationNotification	DELETE
GetDeviceDistance	GET

III. LOCATION DATA RESOURCE STRUCTURE

Location information is presented by latitude, longitude, altitude, accuracy and time stamp. Latitude, longitude and altitude values are expressed as floating point numbers. The accuracy values express the desire of the application for the location information to be provided. The choice of values may influence the price that the service provider would charge. In triggered notifications, a tracking accuracy is defined. Two accuracy values (requested and accepted) may be used. For example, a taxi tracking service that locates the nearest taxi to the client requires fine grained accuracy while coarse grained

accuracy may be appropriate for a truck nearing the vicinity of a warehouse. The accuracy of location provided in meters is expressed as an integer number. In some applications, the maximum age of location information may be useful, e.g. the location information may be cached rather than directly accessed. The maximum acceptance age, in seconds, is expressed in integers.

We followed the ETSI resource structure defined in TS 102690 in modeling location information [16].

Fig.4 shows the resource structure for device location. The deviceLatitude, deviceLongitude and deviceAltitude attributes represent the measured device location and the accuracy attribute represents the measurement accuracy. The timestamp attribute represents the date and time that location was collected. The <subscriptions> sub-resource of the deviceLocation resource contains a collection of 0..n <subscription> resources which represent active subscriptions to location information. Each <subscription> resource has requestedAccuracy and acceptedAccuracy attributes. The requestedAccuracy express the range in which the subscribed application wants to receive location information. The acceptedAccuracy expresses the range that the subscribed application considers to be feasible. If the location cannot be provided within this range, the application prefers not to receive the information. The <deviceDistances>, <devicePeriodicReporting> and <deviceChangeReporting> are sub-resources of the <deviceLocation> resource.

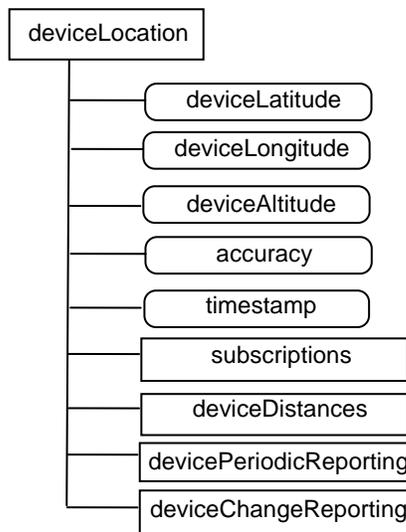


Fig.4 Structure of deviceLocation resource

Some applications may be interested in device distance from a location. The <deviceDistances> collection resource represents the collection of <deviceDistanceFrom> resources. The <deviceDistanceFrom> resource structure is shown in Fig.5, where the remoteLatitude and remoteLongitude attributes represent the latitude and longitude of the location to measure from, respectively. The distance attribute represents the distance from device to the location specified in meters. The subscriptions sub-resource of the <deviceDistances>

resource contains a collection of 0..n <subscription> resources which represent active subscriptions to device distance from the specified location.

Notifications of device location may be provided on a periodic basis. The periodic notifications provide location information at an application defined interval. The <devicePeriodicReporting> collection resource represents the collection of <devicePeriodicLocation> resources. Fig.6 shows the <devicePeriodicLocation> resource structure.

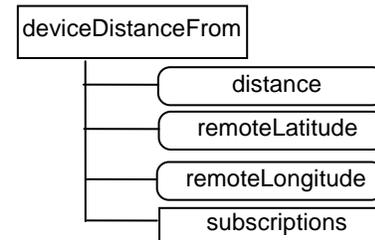


Fig.5 Structure of deviceDistanceFrom resource

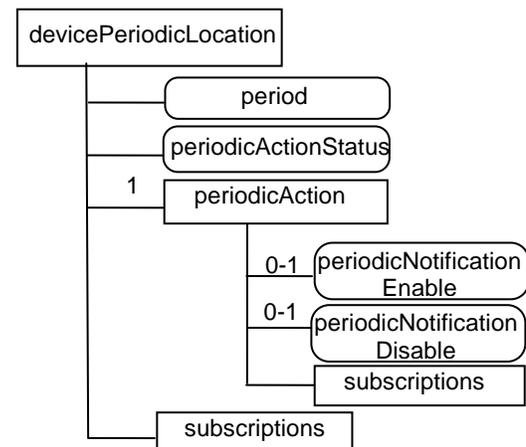


Fig.6 Structure of devicePeriodicLocation resource

The period attribute represents the minimum time between notifications (maximum period of notifications can also be considered). The periodicActionStatus attribute indicates the status of the action. The periodicAction resource represents the action. The periodicNotificationEnable attribute represents the action that enables the periodic notification. The periodicNotificationDisable attribute represents the action that disables the periodic notification.

An application can be notified of a device entering or leaving a geographical area. When a matching event occurs, a notification message will be sent to the application. An application may define a target area and notification criteria e.g. entering the target area or leaving the target area or both. The <deviceChangeReporting> collection resource represents the collection of <deviceLocationChange> resource. The <deviceLocationChange> resource structure for triggered location change notifications is shown in Fig.7. The locationChangeCriteria attribute is of enumerated type (entering or leaving an area). The areaLatitude and areaLongitude attributes represent the latitude and longitude of

the center point, and radius attribute represents the radius of the circle around the center point in meters. The triggeredActionStatus attribute indicates the status of the action. The <triggeredAction> resource represents the action.

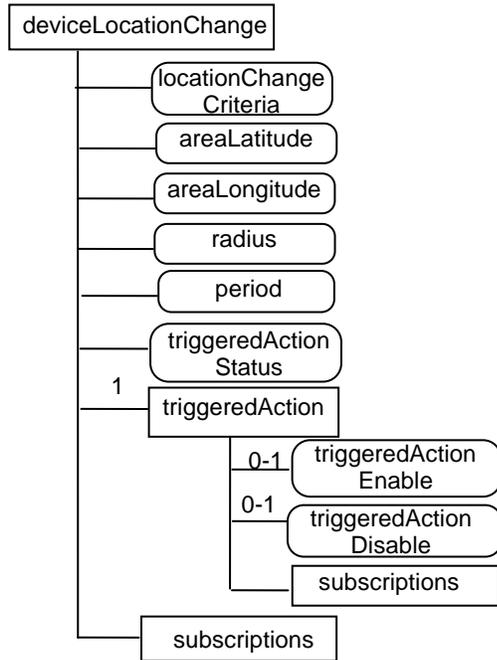


Fig.7 Structure of deviceLocationChange resource

IV. SERVICE CAPABILITY SERVER MODEL

The components used for access rate restriction are of two types, namely Token Bucket mechanism and message filtering. The well known Token Bucket mechanism uses tokens to grant access where tokens arrive at constant rate and are distributed among arriving requests. The message filtering is used to recognize a request in order to apply different access control for CREATE, RETRIEVE, PUT and DELETE requests as specified in SLA. The SCS state is represented by the number of requests that the accepted request queue holds until being forwarded.

Network M2M applications are hosted by service providers. The M2M service capability provider deploys an SCS. The contract between a service provider and the M2M service capability provider defines constraints that have to be fulfilled. The constraints include the peak and average number of network application requests and M2M device notifications that should be accepted per time unit. Access control is applied to each service provider.

Fig.8 shows the SCS access control model built from message filtering (F_i) and token bucket (B_i) components. Each bucket $B_i(T, \rho, \mu)$ is described by the upper limit of tokens T , the token arrival rate ρ_i and the current number of tokens μ_i . The simplest possible model for message filter is the one that checks whether given message belongs to a class. Assuming that for given interval $(t_{k-1}, t_k]$ the flow of messages at the filter input is $N(t_k)$ and the class to filter for is set to be c then

$$F_c(t_k) = \sum_{\forall m \in N(t_k)}^n I(m \in c) \tag{1}$$

where $I(s)$ is 1 when s is true, and 0 otherwise.

It is trivial that the part of the flow filtered out is

$$\widehat{F}_c(t_k) = N(t_k) - F_c(t_k) \tag{2}$$

The queue model is described with respect to the possible loss introduced by its finite length Q . So, it is normal to assume that the initial queue state is empty i.e. $q(t_0) = 0$, and in case it is full, then the queue state becomes $q(t_k) = Q$. Then the loss at the queue for given interval $(t_{k-1}, t_k]$ is formed by all the messages from the arriving flow that come when queue happens to be full, and this might be expressed by

$$L(t_k) = \sum_{t_{k-1} < \tau_i \leq t_k} I(m(\tau_i) \in A(t_k)) \cdot I(q(\tau_i) \equiv Q) \tag{3}$$

The incoming request flow is generated by M2M network applications (U) hosted by a service provider and Device/Gateway M2M applications (V). The granted and the rejected requests are denoted by G and R respectively. The first token bucket B_1 limits the total number of requests. The request is rejected due to violation of SLA, an HTTP 429 *Too many requests* response is sent. The filter F_1 passes DELETE requests all of which are admitted as they are used to release resources. The filter F_2 passes GET requests which are then limited by the token bucket B_2 . The filter F_3 passes PUT requests which are then limited by the token bucket B_3 . Finally, the filter F_4 passes POST requests which are then limited by the token bucket B_4 .

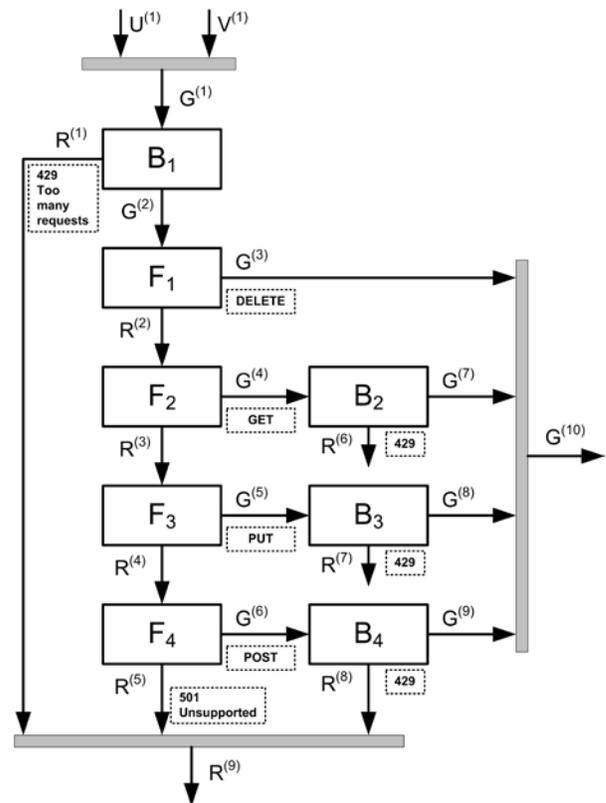


Fig.8 A model of SCS access controller

The request flow is observed in regular intervals $(t_{k-1}, t_k]$. The flows are expressed by the following equations:

$$G^{(1)}(t_k) = U^{(1)}(t_k) + V^{(1)}(t_k) \quad (4)$$

$$G^{(2)}(t_k) = \min(\mu_1(t_{k-1}) + \min(T_1 - \mu_1(t_{k-1}), \rho_1 \Delta t_k), G^{(1)}(t_k)) \quad (5)$$

$$R^{(1)}(t_k) = G^{(1)}(t_k) - G^{(2)}(t_k) \quad (6)$$

$$G^{(3)}(t_k) = \sum_{\forall m_i \in G^{(2)}(t_k)} I(m_i \in c_1) \quad (7)$$

$$R^{(2)}(t_k) = G^{(2)}(t_k) - G^{(3)}(t_k) \quad (12)$$

$$G^{(4)}(t_k) = \sum_{\forall m_i \in R^{(2)}(t_k)} I(m_i \in c_2) \quad (8)$$

$$R^{(3)}(t_k) = R^{(2)}(t_k) - G^{(4)}(t_k) \quad (9)$$

$$G^{(5)}(t_k) = \sum_{\forall m_i \in R^{(3)}(t_k)} I(m_i \in c_3) \quad (10)$$

$$R^{(4)}(t_k) = R^{(3)}(t_k) - G^{(5)}(t_k) \quad (11)$$

$$G^{(6)}(t_k) = \sum_{\forall m_i \in R^{(4)}(t_k)} I(m_i \in c_4) \quad (12)$$

$$R^{(5)}(t_k) = R^{(4)}(t_k) - G^{(6)}(t_k) \quad (13)$$

$$G^{(7)}(t_k) = \min(\mu_2(t_{k-1}) + \min(T_2 - \mu_2(t_{k-1}), \rho_2 \Delta t_k), G^{(4)}(t_k)) \quad (14)$$

$$R^{(6)}(t_k) = G^{(4)}(t_k) - G^{(7)}(t_k) \quad (15)$$

$$G^{(8)}(t_k) = \min(\mu_3(t_{k-1}) + \min(T_3 - \mu_3(t_{k-1}), \rho_3 \Delta t_k), G^{(5)}(t_k)) \quad (16)$$

$$R^{(7)}(t_k) = G^{(5)}(t_k) - G^{(8)}(t_k) \quad (17)$$

$$G^{(9)}(t_k) = \min(\mu_4(t_{k-1}) + \min(T_4 - \mu_4(t_{k-1}), \rho_4 \Delta t_k), G^{(6)}(t_k)) \quad (18)$$

$$R^{(8)}(t_k) = G^{(6)}(t_k) - G^{(9)}(t_k) \quad (19)$$

$$G^{(10)}(t_k) = G^{(3)}(t_k) + G^{(7)}(t_k) + G^{(8)}(t_k) + G^{(9)}(t_k) \quad (20)$$

$$R^{(9)}(t_k) = R^{(1)}(t_k) + R^{(5)}(t_k) + R^{(8)}(t_k) \quad (21)$$

$$R^{(10)}(t_k) = \sum_{t_{k-1} < \tau_i \leq t_k} I(m(\tau_i) \in G^{(10)}(t_k)) \cdot I(q(\tau_i) \equiv Q) \quad (22)$$

The static approach of rate-conditioning is stable but restrictive and this becomes obvious especially in case of event-driven notification scheme when considerable amount of notifications might get discarded while the primary bucket contains tokens.

A partial relaxation of the restriction might be introduced by intra-SLA short-term rate redistribution between secondary buckets. Assuming that the arrival process A is at least short-term stationary, the simplest form of rate redistribution is

$$\rho^{(x)}(t_{k+1}) = \rho_g^{(x)} + \Delta\rho \quad (23)$$

$$\rho^{(x)}(t_{k+1}) = \rho_g^{(x)} - \Delta\rho \quad (24)$$

where x is the second level bucket index subject to $\max_j R_j$,

ρ_g is static guaranteed rate per request type, and y is the bucket index subject to $S_y = \mu(y) + \rho_g^{(y)} \Delta t_k - T^{(y)} - A^{(y)}$ being a positive maximum. If it exists, with no change of total token amount for the amended SLA, the donor's contribution becomes

$$\Delta\rho = \rho_g^{(y)} - \frac{T^{(y)} + A^{(y)} - \mu^{(y)}(t_k)}{\Delta t_k} \quad (25)$$

and it is re-evaluated before each time unit.

The access control is applied for each service provider j . Let us denote by C the SCS's capacity and by N the number of service providers. Then the SCS utilization is

$$\eta = \frac{1}{C \cdot (t_k - t_0)} \sum_{j=1}^N \sum_{i=1}^K (G_j^{(1)}(t_i) - (R_j^{(9)}(t_i) + R_j^{(10)}(t_i))) \quad (26)$$

where K defines the integral time period.

V. SIMULATION RESULTS

The simulation is done on simplified M2M SCS model with three classes of requests. The parameterization of the simulation model is provided by a mobile operator. The capacity of the M2M SCS is 800 requests per second. The behavior of each M2M service provider is modeled by Markov Modulated Poisson Process (MMPP) [17], [18]. New application requests are generated according to four-state MMPP. Changes between different states are uniformly distributed and occur according to Poisson process with mean 4s. The time intervals between POST requests (subscriptions to location notifications), between PUT requests (location notifications) and between GET requests (location retrievals) are exponentially distributed as the arrival process in the context of Web Services with mean 200 s, 50 s, 50 s respectively. The token rate is equal to the guaranteed rate and the bucket size is determined by the peak rate. Initially, $\mu_i(t_0) = T_0$. The length of the interval for observation $(t_{k-1}, t_k]$ is set to 100 ms. The mean processing time for a single request/response is 5 ms.

The aim of simulation is to evaluate the M2M SCS utilization setting different values of guaranteed rates and fixed peak rates. The guaranteed rates for a given M2M service provider define the constraints for preventing the M2M SCS from overloading (GR_1), for the rate of POST requests (GR_2), the rate for PUT requests (GR_3), and the rate of GET (GR_4). The processing capacity of the M2M SCS must be distributed between different types of messages, where the overall message peak rate (PR_1) must be spread between POST requests (PR_2), PUT requests (PR_3), GET requests (PR_4), and DELETE requests.

The simulation is run in a space of SLAs. The SLA $_j$ for j -th M2M service provider consists of the tuple $(GR_1^j, GR_2^j, GR_3^j, GR_4^j)$ for every token bucket of j -th access controller. The values of peak rates $PR_1=100$, $PR_2=12$, $PR_3=38$, $PR_4=38$ are limited by the M2M SCS capacity.

Fig.9 summarizes the outcome of the simulation. The M2M SCS utilization is evaluated as a function of the number of M2M service providers and guaranteed rates.

The simulation results show that the M2M SCS utilization depends both on the number of M2M service providers and on the specific values of rates in SLAs. In case the congestion threshold value is set to 80%, it is most likely that the

appropriate choice is to have 22 M2M service providers applying second type or third type of SLA.

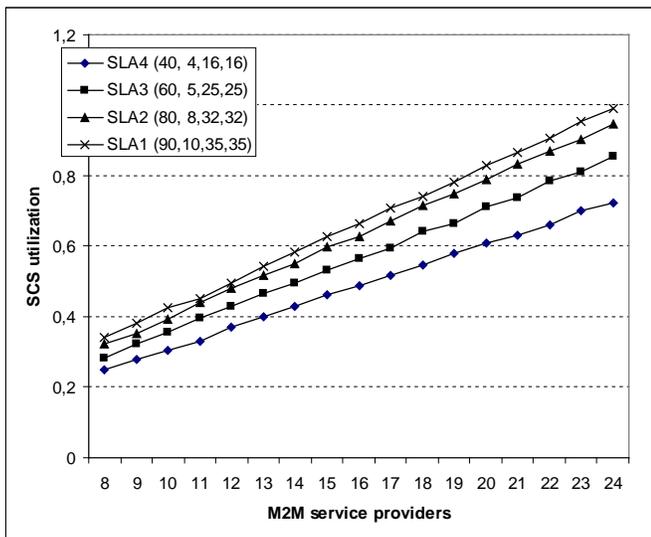


Fig.9 Utilization of M2M SCS versus the number of M2M service providers

VI. CONCLUSION

The paper studies functionality required to support M2M service capabilities that provide access to location information of M2M devices. It may be stated that the contribution of the presented paper is mainly consisted of definition of generic M2M mobility functionality that may be reused across different applications through application programming interfaces. The APIs are based on the REST architectural style.

Different use case scenarios are considered in order to identify basic RESTful Web Service operations. These operations are mapped onto HTTP requests. The semantic information related to mobility is synthesized by identification of basic use cases. The location data is presented in a resource tree structure, where resources may be manipulated through their states. As each of the resources is uniquely addressable, it can be accessed using standard HTTP methods or CoAP (Constrained Application Protocol) primitives.

Deployment aspects of the designed RESTful WS are considered. By evaluation of the SCS load, the values of quality of service related parameters in SLA contracts between M2M service providers and the provider of M2M service capabilities are determined.

The proposed approach is aimed to deal with the fragmented market and it is a step toward development of horizontal M2M platforms.

Our future work concerning development of proposed RESTful Web Services is aimed to be in the field of e-health. In order to accelerate the process of development, it is planned that prototyping of the application will be based on Raspberry Pi like in [19].

REFERENCES

- [1] C. Pereora, A. Aguiar, "Towards Efficient Mobile M2M Communications: Survey and Open Challenges", *Sensors*, vol 14, pp.19582-19608; doi:10.3390/s141019582.
- [2] S. K. Datta, C. Bonnet, "Smart M2M Gateway Based Architecture for M2M Device and Endpoint Management", Available: <http://www.Eurecom.Fr/Fr/Publication/4318/Download/Cm-Publi-4318.Pdf>, 2014.
- [3] J. Kim, J. Lee, J. Kim, J. Yun, "M2M Service Platforms: Survey, Issues, and Enabling Technologies", *IEEE Communications Surveys & Tutorials*, vol.16, no.1, pp. 61-76, pp.2014.
- [4] N. A. Surobhi, A. Jamalipour, "M2M-Based Service Coverage For Mobile Users In Post-Emergency Environments," *IEEE Transactions On Vehicular Technology*, vol. 63, no. 7, pp. 3294-3303, 2014.
- [5] S. Wahle, T. Magedanz, F. Schulze, "Demonstration of OpenMTC – M2M Solutions for Smart Cities and the Internet of Things", Available: http://www.Ieecn.Org/Prior/LCN37/Lcn37demos/Lcndemos12_Wahle.Pdf, 2013.
- [6] F. Bai, K. S. Munasinghe, A. Jamalipour, "A Novel Information Acquisition Technique For Mobile-Assisted Wireless Sensor Networks," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 4, pp.1752-1761, 2012.
- [7] J. Kim, J. Lee, J. Kim, J. Yun, "M2M Service Platforms: Survey, Issues, and Enabling Technologies" *IEEE Communications Survey and Tutorials*, 2014, vol.16, pp. 61–76.
- [8] Y. C. Chang, "Study of Overload Control problem for Intelligent LTE M2M Communication System", *Advances in Smart Systems Research*, vol.3, no.3, pp.44-48, isrp13-004, 2013.
- [9] U. Phuyal, A. T. Koc, M. H. Fong, R. Vannithamby, "Controlling Access Overload and Signaling Congestion in M2M Networks", presented at ASILOMAR'2012 Conference on Signals, Systems and Computers, Conference proceedings, pp.591-595, 2012.
- [10] S. Duan, V. S. Mansouri, V. Wong, "Dynamic Access Class Barring for M2M Communications in LTE Networks", presented at Wireless Networking Symposium, Globecom'2013, pp.4747-4753, 2013
- [11] T. Taleb, A. Ksentini, A. Kobbane, "Lightweight Mobile Core Networks for machine Type Communications", *IEEE Access*, 2014, vol.2 pp.1128-1137.
- [12] Y. J. Chen, Y.H. Shen, L.C. Wang, "Traffic aware Load Balancing for M2M Networks Using SDN", presented at IEEE International Conference on Cloud Computing technology and Science, 2014, Conference proceedings, pp.668-671.
- [13] S. K. Datta, C. Bonnet, N. Nikaiein, "An IoT Gateway Centric Architecture to Provide Novel M2M Services", presented at IEEE World Forum on Internet of Things, WF-IoT'2014, Conference proceedings, pp.514-519, 2014.
- [14] J. Latvakoski, A. Iivari, P. Vitic, B. Juben, M. Alaya, T. Monteil, Y. Lopez, G. Talavera, J. Gonzalez, N. Granqvist, M. Kellil, H. Ganem, T. Vaisanen, "A survey on M2M Service Networks", *Computers*, vol.2, pp.130-173, 2014.
- [15] R. Ratasuk, A. Prasad, Z. Li, A. Ghosh, M. Uusitalo, "Recent Advancements in M2M Communications in 4G Networks and Evolution Towards 5G", presented at International Conference on Intelligence in Next generation Networks, ICIN'2015, Conference proceedings, pp.1.-6, 2015.
- [16] ETSI TS 102 690, "Machine-to-Machine communications (M2M); Functional architecture", 2011.
- [17] J. Tang, W. P. Tay, Y. Wen. "Dynamic Request Redirection and Elastic Service Scaling in Cloud-Centric Media Networks", *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1434-1445, 2014.
- [18] J. Latvakoski, M.B. Alaya, H. Ganem, B. Jubeh, A. Iivari, J. Leguay, J. M. Bosch, N. Granqvist, "Towards Horizontal Architecture for Autonomic M2M Service Networks", *Future Internet*, vol.6, pp.261-301, 2014.
- [19] G. Stoyanov, B. Naidenov, S. Kostadinova. "Using of mobile platforms for sensor nodes in Biomedical Wireless Sensor Networks", 50th International Scientific Conference on Information, Communication and Energy Systems and Technologies ICEST'2015, Preliminary Proceedings, 2015.

Performance Estimation of Non-comparison Based Sorting Algorithms Under Different Platforms and Environments

Mentor Hamiti and Diellza Nagavci

Abstract—There may be several different platforms for performance estimation of non-comparison based sorting algorithms. Understanding the relative efficiencies of algorithms designed to do the same task is very important in every area of computing. An algorithm can be analyzed in terms of time efficiency or space utilization. The efficiency, with which a sorting will be carried out, often has a big impact on the effectiveness of the program as a whole. Because platforms and surroundings are in progression and permanent change, they always need to follow these parameters. The goal of this paper is to review different non-comparison based sorting algorithms. In this occasion three different environments and computer performances are used and the obtained results are also analyzed in this paper.

Keywords—algorithm, environment, non-comparison, platform, sorting.

I. INTRODUCTION

SORTING is maybe the single most important algorithm performed by computers, and certainly one of the most investigated topics in algorithmic design. One of the fundamental problems of computer science is ordering a list of items. There is a plethora of solutions to this problem, known as sorting algorithms. An algorithm can be analyzed in terms of time efficiency or space utilization. The running time of an algorithm is influenced by several factors: speed of the machine [2] running the program and language in which the program was written; Efficiency of the compiler that created the program, the size of the input and the organization of the input. Examples of sorting algorithms that run in linear time are counting sort, radix sort and bucket sort are executed in three platforms as CPU: Intel® Core i5™-M460 2.53GHz (2 Cores), RAM: 6GB, CPU: Intel® Core i3™-2100 3.10GHz (2 Cores), RAM: 4GB and also in Pentium® Dual Core – T4200 2.0GHz (2 Cores), RAM: 4GB and three environs as C++, Python and Java.

II. NON-COMPARISON SORT

A. Bucket Sort

Bucket Sort is a sorting method that subdivides the given data into various buckets depending on certain characteristic

order, thus partially sorting them in the first go. Then depending on the number of entities in each bucket, it employs either bucket sort again or some other ad hoc sort. Bucket sort runs in linear time on an average. Bucket sort is stable. It assumes that the input is generated by a random process that distributes elements uniformly over the interval 1 to m . Bucket sorting algorithm is a kind of sustainable, [1] it takes data generated by a random process that distributes the same elements in the interval $O(n)$. Bucket sort divides the intervals $[0,1)$ in the same size intervals or bucket and then distributes them in the data bucket. Once the data are distributed uniformly and in the interval $[0,1)$ we do not expect that each number will enter the empty bucket-mails. To gain done sorting scoring numbers in each bucket and then go to the order of bucket's listed the elements in the list.

B. Counting Sort

Counting sort is an algorithm used to sort data whose range is pre-specified and multiple occurrences of the data are encountered. It is possibly the simplest sorting algorithm. The essential requirement is that the range of the data set from which the elements to be sorted are drawn is small, compared to the size of the data set [3]. Counting sort works by determining how many integers are behind each integer in the input array A. Using this information, the input integer can be directly placed in the output array B. This type of sorting works best when data distribution is uniform. An example of efficient use of Counting Sort order can be 200 students on the basis of their results by sorting 100 or 1500 employees in connection with the filing of their birthday in a year. The drawback may occur if range $m \gg n$ (where n is the number of data while m is the range of data), the complexity will not be linear in n and thus this sort will not remain useful longer. This is because the chances of the appearance of gaps, during the sorting for those elements which do not exist in the list will cause a higher complexity of space. Because counting sort algorithm is a straightforward algorithm is quite simple and easy to be analyzed in the context of software complexity. The worst case and the average performance of counting sort algorithm is $O(n + k)$. In order to ensure maximum efficiency, "k" should not be higher than "n". Counting Sort When compared with other sorting algorithms, appears to be easier to

implement and does not require any special structure of data to store its elements.

C. Radix Sort

A radix sort is an algorithm that can rearrange integer representations based on the processing of individual digits in such a way that the integer representations are eventually in either ascending or descending order. Integer representations can be used to represent things such as strings of characters (names of people, places, things, the words and characters, dates, etc.) and floating point numbers as well as integers. So, anything which can be represented as an ordered sequence of integer representations can be rearranged to be in order by a radix sort [4]. Most digital computers internally represent all of their data as electronic representations of binary numbers, so processing the digits of integer representations by groups of binary digit representations is most convenient. Two classifications of radix sorts are:

- Least significant digit (LSD) radix sort.
- Most significant digit (MSD) radix sort.

LSD radix sorts process the integer representations starting from the least significant digit and move the processing towards the most significant digit. MSD radix sorts process the integer representations starting from the most significant digit and move the processing towards the least significant digit. The integer representations that are processed by sorting algorithms are often called "keys," which can exist all by themselves or be associated with other data. LSD radix sorts typically use the following sorting order: short keys come before longer keys, and keys of the same length are sorted lexicographically. This coincides with the normal order of integer representations, such as the sequence 1, 2, 4, 5, 6, 7, 8, 9. MSD radix sorts use lexicographic order, which is suitable for sorting strings, such as words, or fixed-length integer representations. A sequence such as b, c, d, e, g, h, i, j, ba would be lexicographically sorted as b, ba, c, d, e, f, g, h, i, j.

If lexicographic ordering is used to sort variable-length integer representations, then the representations of the numbers from 1 to 10 would be output as 1, 10, 2, 3, 4, 5, 6, 7, 8, 9, as if the shorter keys were left-justified and padded on the right with blank characters to make the shorter keys as long as the longest key for the purpose of determining sorted order.

III. TESTING IN DIFFERENT PLATFORMS AND ENVIRONMENTS

The three non-comparison algorithms that are tested for different number of CPUs will enable finding the best ratio of the volume of data to the number of cores. For example Bucket sort is implemented in three platforms, Radix Sort and Counting Sort in three platforms and in C++, Python and Java environments.

A. Bucket sort implementation CPU platform: Intel® Core i5 (TM) 2.53GHz, 6GB RAM in C++ environment

To analyze an algorithm we should provide tools which are used by an algorithm for functioning. In the general case these

tools are: space memory devices, generation communications or computer hardware and execution time. Bucket sort algorithm is implemented in C++ environment executed in Visual Studio 2013.

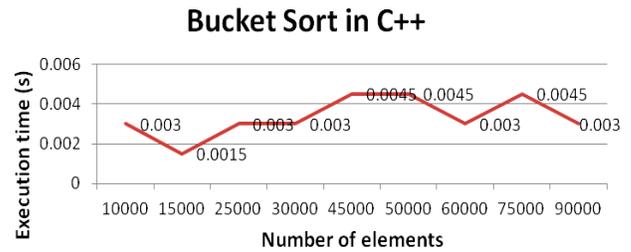


Fig. 1. Results of Bucket sort execution time

The diagram shows bucket sort execution time by the number of elements. With increasing the size and the number also increases the execution time on this platform. The best time execution is 0.003 seconds.

B. Bucket sort implementation CPU platform: Intel® Core i5 (TM) 2.53GHz, 6GB RAM in Python environment

Diagram for bucket sort in python environment presents results that show the curve through the highest point of the execution time in this case is thus 0.065 seconds in the range of 10000 to 100000 numbers.

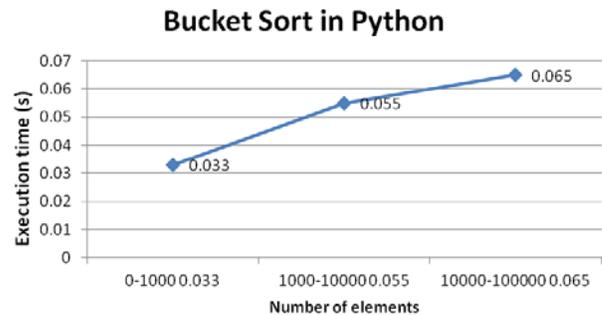


Fig. 2. Results of Bucket sort execution time in Python

C. Radix sort implementation CPU platform: Intel® Core i5 (TM) 2.53GHz, 6GB RAM in Java environment

Speed of radix sort largely depends on the inner basic operations and if operations are not efficient enough radix sort can be slower than some other algorithms such as quick sort or merge sort.

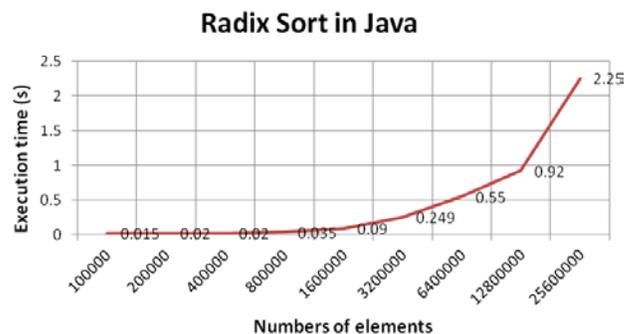


Fig. 3. Radix Sort in different platforms in Java environment

These operations include the insert delete function of the sub lists and the process of isolating the digit we want.

Based on this graphic we can conclude that radix sort in 25600000 elements had a worst case of exestuation, otherwise the best case is 0.015 seconds in 100000 elements.

D. Counting sort implementation CPU platform: Intel® Core i3 (TM)2100 3.10 GHz, 4GB RAM in Python environment.

Counting sort is implemented in Python environment, this non-comparison algorithm is stable. The best case of exestuation time is 0.009 seconds.

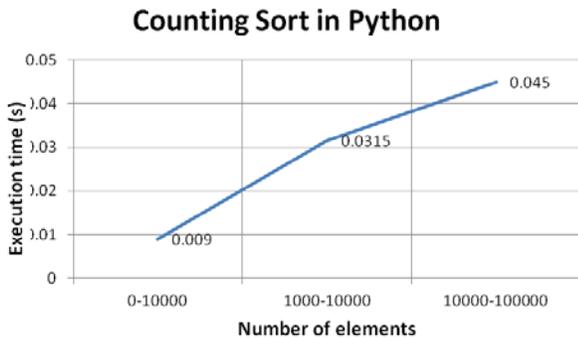


Fig.4. Counting Sort in different platforms in Python environment

E. Radix sort implementation CPU platform: Intel® Core i3 (TM)2100 3.10 GHz, 4GB RAM in Java environment.

Implementation of Radix Sort in Java environment with CPU platform: Intel® Core i3, has different results, the best case is 0.022 seconds. The Java was compiled in Eclipse.

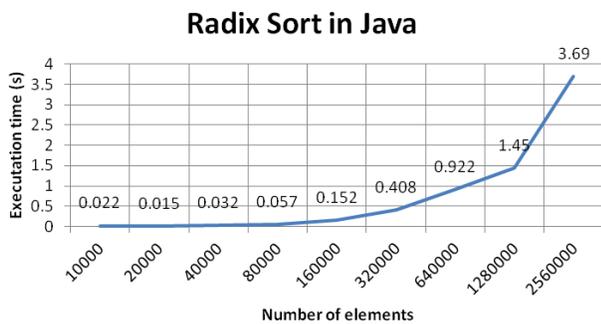


Fig. 5. Radix Sort in different platforms in Java environment

F. Bucket sort implementation CPU platform: Pentium® Dual Core 2GHz, 4GB RAM in C++ environment.

Bucket sort algorithm is implemented in platform CPU: Pentium® Dual Core 2GHz, in the C++ environment compiled in Visual Studio 2013. As we can see in the figure the worst exestuation time is 27.852 seconds, but the best case is 0.434 seconds. We can conclude that the execution time depends in the number of elements.

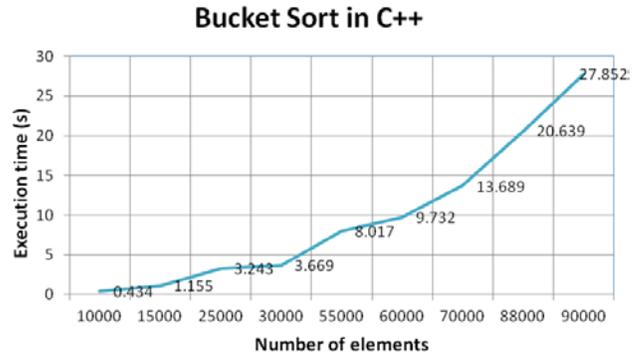


Fig. 6. Bucket Sort in different platforms in C++ environment

G. Radix sort implementation CPU platform: Pentium® Dual Core 2GHz, 4GB RAM in Java environment.

Radix Sort is implemented in java environment and in platform with Pentium® Dual Core 2GHz. The code is compiled in Eclipse. Based in the figure we can conclude that the best case of execution is 0.02, increasing the number of elements the execution time will score 12.25 seconds.

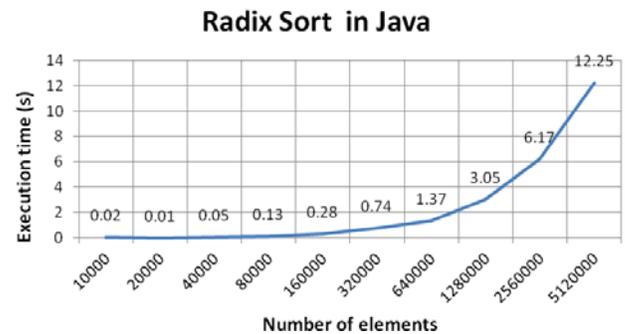


Fig. 7. Radix Sort in different platforms in Java environment

IV. ANALYSIS OF THE RESULTS OBTAINED DIFFERENT PLATFORMS AND ENVIRONMENTS

Non-comparison algorithm Bucket, Radix and Counting sort are tested in configurations with various performance and by that we conclude which algorithm is executed most quickly. Also we can reach a point where we see the results obtained by each algorithm without comparisons. These three algorithms implemented in three programming languages selected for this study will serve as a benchmark of the results obtained by different configurations, including computers with processors i5, i3 and Pentium dual-core.

A. Bucket Sort in different platforms in C++ environment

As we can in the Figure Bucket sort algorithm is implemented in three different platforms featuring distinction at the time of execution, depending on the characteristics of the computer.

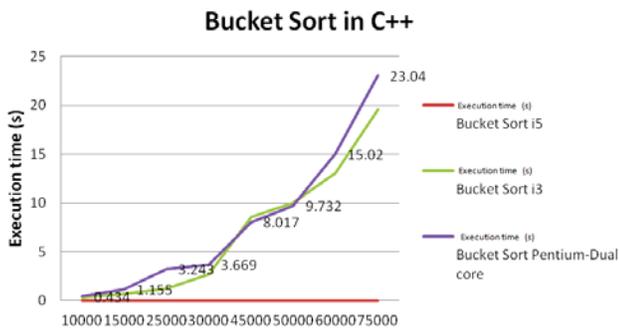


Fig. 8. Bucket Sort in different platforms in C++ environment

B. Bucket Sort in different platforms in Python environment

As in the previous case, we conclude that the worst implemented algorithm is in the Python programming language in surroundings Pentium Dual Core where the duration of the performance is 68.78 seconds against times significantly faster computers with processors i3 and i5. Below is the report in tabular and graphical form.

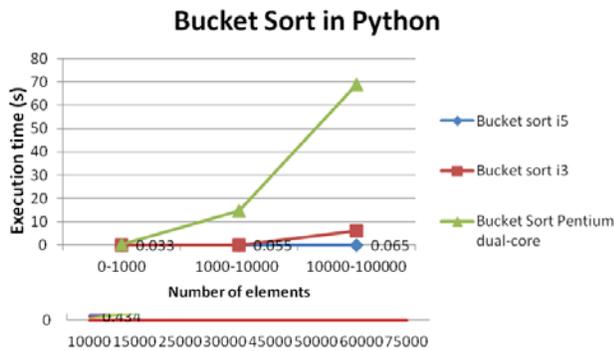


Fig. 9. Bucket Sort in different platforms in Python environment

C. Bucket Sort in different platforms in Java environment

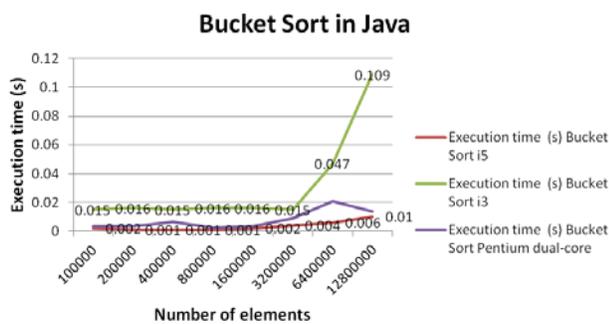


Fig. 10. Bucket Sort in different platforms in Java environment

In the diagram above we can see the results showing the best case and worst case. As best case is the platform with i5 processor which for a short time e executes all data elements in the Java programming language and as a worst case according to the results is the performance of the computer with Pentium Dual Core processor. The best case then algorithm execution is i5 processor.

D. Radix Sort in different platforms in C++ environment

Radix sort algorithm is implemented in the vicinity of compiled C ++ in Visual Studio 2013 in various performance processors i5, i3, and Pentium dual-core. Based on the results that are obtained we can conclude which Radix sort algorithm performance is better and which worse. Below is the results of comparisons between different performances of Radix sort algorithm.

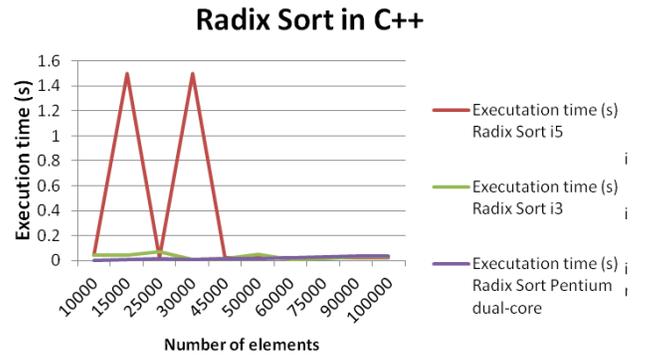


Fig. 11. Radix Sort in different platforms in C++ environment

E. Radix Sort in different platforms in Python environment

Radix sort algorithm is implemented in the environment of the compiled Python “Python GUI” in various performance i5, i3, and Pentium dual-core. Based on the obtained results we can conclude for the best and worst performance of Radix sort algorithm. Below are the results of comparisons between different performances of Radix sort algorithm.

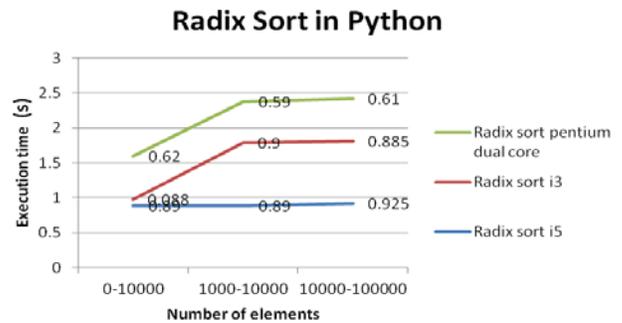


Fig. 12. Radix Sort in different platforms in Python environment

F. Bucket Sort in different platforms in Java environment

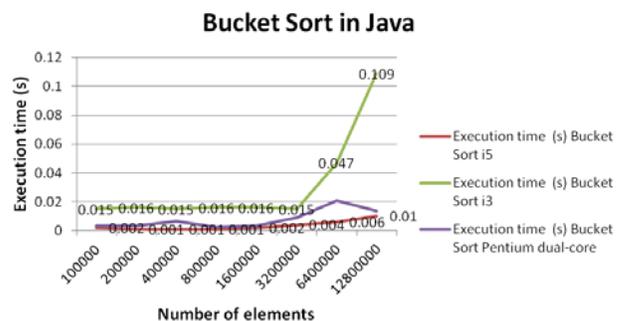


Fig. 13. Bucket Sort in different platforms in Java environment

Bucket sort algorithm is implemented in three different platforms i5 processor performance, and Dual-Core i3. Based on these results it was concluded which of these algorithms is most appropriate for the respective performance.

G. Counting Sort in different platforms in C++ environment

Counting sort algorithm is implemented in C ++ environment, which is compiled in Visual Studio 2013 in three different platforms i5 processor performance, and Dual-Core i3. Based on these results it was concluded which of these algorithms is most appropriate for the respective performance.

Counting Sort in C++

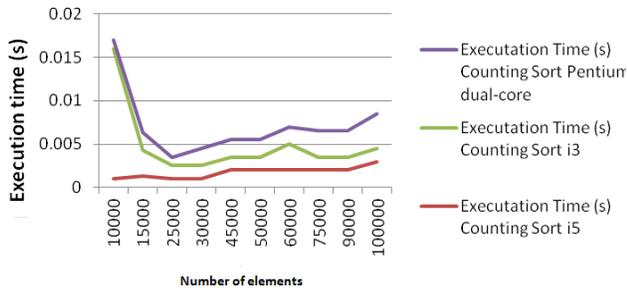


Fig. 14. Counting Sort in different platforms in C++ environment

H. Counting Sort in different platforms in Python environment

Counting sort algorithm is implemented in Python environment which is compiled on the Python GUI and on three different platforms i5 processor performance, and Dual-Core i3. Based on these results it was concluded which of these algorithms is most appropriate for the respective performance.

Counting Sort in Python

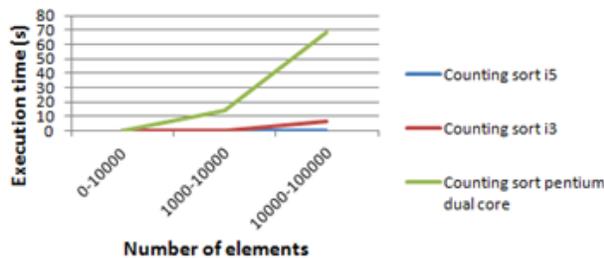


Fig. 15. Counting Sort in different platforms in Python environment

I. Counting Sort in different platforms in Java environment

Counting sort algorithm is implemented in Java environment which is compiled in Eclipse in three different platforms i5 processor performance, and Dual-Core i3. Based on these results it was concluded which of these algorithms is most appropriate for the respective performance.

Counting in java

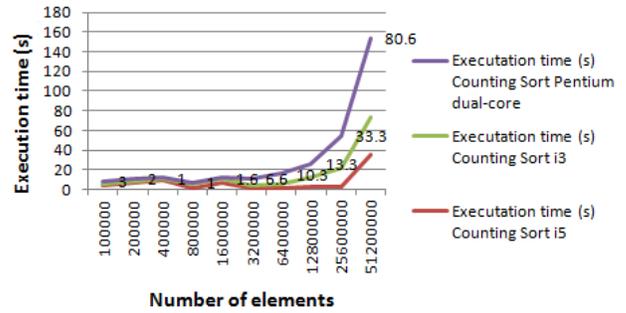


Fig. 16. Counting Sort in different platforms in Java environment

V. CONCLUSION

In this paper, we have studied and analyzed about non-comparison based sorting algorithms. We analyzed the time complexity of each algorithm with time taken by each step of algorithm in different platforms and environments.

The main objective to analyze the performance of sorting algorithms without comparisons focused on various localities, including programming languages, such as C ++, Python, Java, by analyzing the programs for taking their time for execution. Initially, the description of the algorithm for ranking, then the assessment of sorting algorithms, sorting algorithms classification without comparisons, which represents the complexity of algorithms without comparison, as variable memory devices it increases the number of data. However the essence of this work has been the performance estimation of non-comparison based sorting algorithms under different platforms and environments specifically the implementation of algorithms Bucket, Counting and Radix Sort platforms with processors i5, i3 and Pentium Dual-Core in various localities such as C ++, Python and Java. From the results obtained we can conclude that the algorithm ran in Python environment Counting sort is the best in i5 platform, followed by sort and bucket Radix sort algorithm as last. The obtained results in i3 platform, shows that the first algorithm for this platform is Counting sort, followed by Radix sort and Bucket sort as last. In Pentium Dual-Core platform best algorithm is Counting sort, second is Radix sort and third is Bucket sort. For Java environment in i5 platform the most appropriate algorithm is Bucket sort, the second is the Counting sort and final is Radix sort. For i3 performance in Java environment as first algorithm is Bucket sort, the second is Counting sort and final remains Radix sort. In Pentium Dual-Core performance in Java environment, the best algorithm for this case is Counting sort, second is Bucket sort and last is Radix sort. The results obtained from the implementation of non-comparison algorithms in different platforms have different results. But what is most important is that non-comparison sorting algorithm has the best rating in C++ environment compared with sequential equivalent solutions on platforms Java and Python. After platforms were tested implementations are achieved where the execution time is faster in C++ environment than in Python and Java environment.

REFERENCES

- [1] Spirakis. (2013). Algorithms and Complexity.
- [2] Introduction to Algorithms by Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Second Edition 2012 (Prentice Hall of India private limited), New Delhi-110001.
- [3] Parag Bhalchandra, Proliferation of Analysis of Algorithms with application to Sorting Algorithms, M.Phil
- [4] Dissertation, VMRF, India, July 2011.
- [5] Mishra, A. B., 2013. Comparison of Sorting Algorithms based on Input Sequences International. Journal of Computer Applications.
- [6] Soltys, M., 2012. An Introduction to the Analysis of Algorithms. s.l.:s.n.
- [7] Mahfooz Alam, A. C., 2014. Sorting Algorithm: An Empirical Analysis. International
- [8] Journal of Engineering Science and Innovative Technology (IJESIT).
- [9] Comparison of Sorting Algorithms based on Input Sequences International Journal of Computer Applications (0975– 8887)Volume 78 – No.14, September 2013.

Reduced Permissions Schema for Malware Detection in Android Smartphones

Ahmed H. Mostafa, Marwa M. A. Elfattah and Aliaa A. A. Youssif

Abstract— Day after day the dependence on smart devices is increasing, especially smart phones. As, smartphone is not just a phone device but also it is smart TV, GPS, smart camera and tablets, with expansion in the use of mobile in critical tasks such as online banking services, business transactions, and storing critical information such as credit cards, passwords and personal data, the malware's attacks are increased. Most of current malware detection solutions for mobile devices can detect known malware but cannot detect newfangled malware and others malware detection techniques depend on monitoring the behavior of the malware but the monitoring on the Smartphone can be a very heavy consuming task.

Hence, there is a need to develop a mobile malware detection that can provide an effective solution to protect the mobile user from any malware and at the same time address the limitation of mobile devices environment. In this paper we focused on extracted android system permissions from android applications .apk files. The research focused in reducing the number of android permissions to be used as features for machine learning classifier to detect android malware application.

Keywords— Android, Smartphones, Malware Detection, Machine learning.

I. INTRODUCTION

Now a day's, smartphone is very popular and widely used in business and personal life. A lot of mobile phone users are rapidly switching to smartphones. According to eMarketer [1], it is expected that around 49% of the mobile phone users globally are likely to use smartphones by 2017. The great popularity of the smartphones is because of their powerful capabilities such as video calling, capturing images, recording video, playing digital media, sending and receiving emails, web browsing and access online banking services capturing images, recording audio and video, video calling, playing digital media, sending and receiving emails, web browsing, using social networks such as Facebook and Twitter, and communicating using Bluetooth and WIFI.

Smartphone users save important data on their phones, such as phone numbers, SMS messages, photos, passwords, credit

card numbers, therefore smart phones are a very interesting target for attackers and malicious software. One important characteristic of smartphones that its ability to install third-party applications from many markets whether official or non-official. Unfortunately, there is no control on the non-official markets, therefore attackers can upload their application whether games, media or others applications to these markets and attempt to embed malicious program into benign applications

Many users download mobile applications without any thought of security. Whereas, with the rapid increase in the use of smartphones, the number of mobile applications is increasing, and according to PortioResearch [2] downloading of mobile applications will continue to grow to exceed 200 billion applications by the end of year 2017, The number of markets which allow users to download applications are increasing and the number of non-official market are also increasing but non-official market do not impose security measures on the phone applications that are being uploaded by developers so many hackers upload malicious applications to these markets

Actually, most smartphones users download mobile applications without any attention of security issue. Therefore, it is important to use a methodology to detect the malware applications before installing it on the phone.

The problem of detecting malware for smartphone presents a lot of challenges due to limited resources availability. Smartphones have limited hardware capabilities in comparison to the hardware capabilities of traditional computers, as smartphones have limited memory, and limited battery energy. So, current solutions for computers may not be applicable on smartphones.

Moreover, most current malware detection techniques depend on extract signatures pattern for malwares, and all malwares signatures are stored in repository. This repository represents malwares signature database to identify malware. The antivirus should search in the database for matching signatures, but it cannot detect new malwares.

On other hand, the other set of techniques depend on monitoring the behavior of malware during the run time but monitoring can be a very heavy consuming task [3]-[5].

Monitoring can be performed on remote servers but it is dependent on external server, which means there can be server down problems and network congestion [6].

With very rapid development in smartphones also its operating system have evolved so the techniques and approaches that applied on the previous Mobile operating

This work was supported by Computer Science Department, Faculty of computers and information, Helwan University.

Ahmed Hesham Mostafa, Computer science Department, Faculty of computer and information, Helwan university, Cairo, Egypt (phone: +2001095906541; e-mail: ahmedheshamostafa@gmail.com).

Dr. Marwa Mohamed Abd El Fatah, Computer science Department, Faculty of computer and information, Helwan university, Cairo, Egypt (e-mail: marwa_8_80@yahoo.com).

Prof. Aliaa Youssif, Computer science Department, Faculty of computer and information, Helwan University, Cairo, Egypt (Phone: +202 27644827; Fax: +202 25547975; e-mail: aliaay@yahoo.com)

system need to be modified to be applicable with current operating system.

The most common mobile operating systems are Android, Blackberry, iOS, Windows Phone and Symbian.

Statista [7] expected that Android is expected to account for 62.4 percent of global tablet shipments in 2017, thus taking over as the market leader. Statista also expected that the smartphones deploying Android as operating system are forecast to reach around 1.5 billion units by 2018[8]. Cisco security report for 2014 finds 99% of all new mobile malware is targeting Android [9]. Android's Google Play store has officially reached over 1 million applications, and applications download have also grown to over 50 billion [10]. Several third-party Android Marketplaces exist without restricted security rules for submit applications.

The challenge of how to detect smartphones malwares depends mainly on how to extract the application features. Those features are then used to categorize the application as malware or as benign application. This research introduces a mechanism to select reduced number of application features, which are used in anomaly detection system to detect android malwares before installing it on smartphone.

The rest of the paper organized as follows. We start in section II with a brief background on malware detection techniques, in section III a survey of previous relevant studies, in section IV brief background on android operating system, in section V describes the methods we used to collect data, extract features and building the dataset, in sections VI, VII we present the experiments and the evaluation results. Finally in section VIII discuss the results and conclusion.

II. MALWARE DETECTION TECHNIQUES

There are two main categories of smartphones malware detection techniques, which are static detection techniques and dynamic detection techniques [3]-[5]. The major difference between static and dynamic analysis is how the data is acquired.

Static detection represents an approach of checking source code or compiled code of applications before it gets executed. It identifies malicious code by unpacking and disassembling the application to extract features for anomaly detection. It can use simple pattern search operation or slightly more complex machine learning approaches in order to detect weakness in the code of software.

On other hand, dynamic set of techniques identify malicious behaviors after executing the application on an emulator or controlled environment.

Static based techniques are fast, flexible and easy to be automated, which means, they are suitable for mobile devices whereas, in dynamic based analysis the monitoring can be a very heavy consuming task. Also, in dynamic based analysis, the malware can change his behavior during rum time and cannot be detected.

On other hand, there is different identification techniques depending on the type of identification carried out, detection systems can be classified as either anomaly-based, signature based system.

Anomaly-based identification attempts to model normal and non-normal behaviors during the training phase. Anomaly detection techniques have the potential to detect newfangled malware. However, they are prone to detect rare legitimate behaviors as malicious [5].

Signature-based identification aims at identifying known malicious by means of predefined patterns of signatures. The main benefit of signature detection lies in its accuracy detecting well-known attacks. In this regard, maintaining an up-to-date database with a massive amount of signatures poses a major challenge. Furthermore, resource-constrained devices are not capable of processing big amount of signatures [5]. Also, they need human expertise to develop new malware signatures, which is time consuming.

Static signature-based technique is very efficient and reliable to identify known malwares; otherwise, they cannot detect unknown malwares. Signatures must be up-to-date that lead to a massive amount of signatures. On other hand, anomaly based techniques have the ability to detect unknown malwares.

III. RELATED WORK

Crowdroid [11] and MADAM [12] are among the research works that perform android malware detection by monitoring the dynamic malware behavior through the system call. The drawback of this method is the high energy consumption as monitoring system calls consume lots of resources of a mobile device. Yerima [13] proposed approach based on Bayesian classification models obtained from static code analysis to detect android malware. Borja Sanz [14] proposed PUMA which they extract permission to train machine learning algorithm and they use all permissions and the best accuracy result with RandomForest is 0.8637. Xing Liu proposed a two-layered permission based detection scheme for detecting malicious Android applications [15].

IV. ANDROID SYSTEM ARCHITECTURE

Android [16] is open source OS built on Linux for mobile devices. As shown in Fig. 1, Android system consists of:

- Linux kernel provides basic system services, such as process scheduling.
- Intermediate layer include Android native libraries and Android runtime environment.
- Android native libraries include core libraries such as the system C library, media libraries; various system components in the upper layers use these libraries.
- Android runtime environment is the Dalvik virtual machine.
- Application Framework layer is which make it easy for developers to develop new applications.
- Application layer include core applications, such as call, message and third-party developed applications.

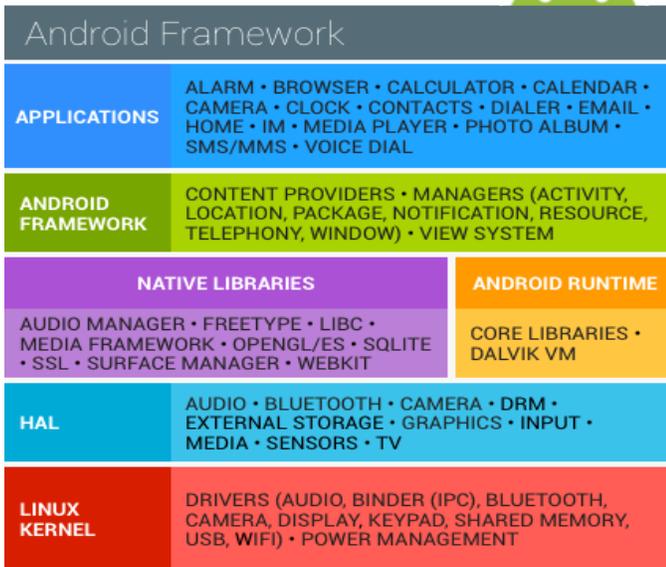


Fig. 1 Android Software Stack [17]

Android applications are developed with Google Android SDK [18] and written in Java language. Then the source code is compiled into .dex file, and packaged in an .apk archive for installation.

Android permits application installation from third party vendors mean that Google has no control over the quality or safety of the applications provided in these stores.

Android Permissions is a critical design point of the Android security architecture is that no application has permission to perform any operation that would impact other applications, the operating system, or the user, this includes reading or writing the user's private data (such as contacts or e-mails), reading or writing another application's files, performing network access etc [19].

Android sandboxes applications from each other so, application must explicitly share resources and data. They do this by declaring the permissions they need for additional capabilities not provided by the basic sandbox. Applications statically declare the permissions they require, and the Android system prompts the user for consent at the time the application is installed [19].

V. RESEARCH METHODS

The analysis of applications is often to classify an application as malicious or benign. In classification features are used to make decisions. Application features are required to be informative to produce an accurate decision.

In many real-world applications, numerous features are used in an attempt to ensure accurate classification. If all those features are used to build up classifiers, then they operate in high dimensions, and the learning process becomes computationally and analytically complicated. Hence, there is a need to reduce the dimensionality of the feature space before classification.

This work mainly aims to extract android application features based on dimensionality reduction technique that extracts a subset of new features from the original set of features by means of some functional mapping keeping as much information in the data as possible.

A. Collect Data

To conduct experiments, a dataset of real Android applications and real malware is considered. In particular, an initial dataset of 325 malware and 325 benign android applications is acquired. The malwares are collected from Contagio Malware Dump [20] Android Malware Dump [21] and MalShare [22].

Malware applications represent more than 89 android malware families [23], [24]. Whereas, the benign applications cover all android categories in Google play store [25].

B. Extract Features

Android permissions control the access to sensitive resources and functionalities. Permissions allow an application to access potentially dangerous API functionality. Many applications require several permissions to function properly. These permissions must be listed explicitly in the application's Manifest.xml file. Every application must have an android Manifest.xml in its root directory. The manifest presents essential information about the application to the Android system.

Using the permissions as features for machine learning classifier can help to detect the malware before the installation. So, analyzing the android applications manifest files to identify the permission set requested by that application can be considered as an informative methodology for anomaly based feature extraction in static manner.

First of all, the application .apk file is decompressed to retrieve the content. All permissions used by each APK file are extracted statically using python [26] script that is developed based on AndroGuard API [27]. We developed a python script to automate the extraction of the features. The developed script unpack the apk files to classes.dex and the manifest file in binary format, then convert the manifest to xml file, where, all permissions used by the application can be extracted.

All permissions from manifest files are extracted based on the following methodology:

- Vector V contains all android system permissions.
- For each application there is features vector V_i contains all features for each application the feature vector represents all android system permissions. So, for each application a_i in the Applications set A there is binary vector $V_i = \{v_1, v_2, v_3, \dots, v_n\}$ where, n is number of permissions available in the Android system, and,

$$v_i = \begin{cases} 1, & \text{if extracted permission } v_n \text{ exist in } V \\ 0, & \text{else} \end{cases}$$

- The variable C is the type of the applications to be benign or malware where $C \in \{Malware, Benign\}$
- The creating of matrix M process is described by following algorithm :

Input: set A contain all apk files and vector V contain all android system permissions

Output: matrix M contain all vectors V_i

for each a_i **in** A **do**

 Extract all permissions from a_i and set it to set S_i

for each $s_j \in S_i$ **do**

if $s_j \in V$ **do**

$v_n \in V_i = 1$

else

$v_n \in V_i = 0$

end if

end for

 Set V_i in M

end for

After applying the previous methodology on all of the collected dataset, we noted, that the benign application use 1141 permissions and malware application use 4882 permissions. Approximately, malware applications use nearly three times permissions more than benign applications, that means malwares actually use permissions to access functions the benign applications not use.

C. Reducing number of features

Actually, we count 151 android system permissions according to android 5.0 Lollipop with API level-21 [28],[29] considering all of android permissions as a feature set will produce an enormous feature vector for each application. So it is required to reduce the number of the application features, where the high dimension data makes testing and training of general classification methods complicated.

The goal of data reduction is to find a minimum set of features such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all features. Using the reduced set of features has additional benefits. It reduces the number of features appearing in the discovered patterns, helping to make the patterns easier to be understood. Further it enhances the classification accuracy and learning runtime.

In the conducted experiments, applying the previously stated methodology for feature extraction based on android permissions produced a matrix M , which contains the vectors of the android system permissions of all collected applications. For reducing the feature set, a preprocessing step has been performed, which is removing all zero-frequency-permissions in the binary matrix M . The permissions that its frequency is

zero are those which are not used by any malware or benign applications, the number of features were reduced to 114 features.

Then, to select the most informative feature set, two feature selections methodologies are applied [30],[31], namely information gain and Gain Ratio based feature selection methods. The information gain and Gain Ratio score are calculated for each permission - attribute in M matrix- that is for telling how important a given attribute of the feature vectors is.

InfoGain and GainRatio are calculated as following:

$$InfoGain(C, v_n) = H(C) - H(C | v_n) \quad (1)$$

$$GainRatio(C, v_n) = (H(C) - H(C | v_n)) / H(v_n) \quad (2)$$

Where H is the information entropy, Y and X are random variables and P is the probability.

$$H(X) = -\sum_i P(x_i) \log_b P(x_i) \quad (3)$$

Where the conditional entropy of two events X and Y

$$H(X | Y) = -\sum_{i,j} P(x_i, y_j) \log \frac{P(y_j)}{P(x_i, y_j)} \quad (4)$$

After calculating the score for each permission, all permissions that its score is zero are removed. Two sets of features (permissions) have been produced, the first set of permissions calculated by InfoGain shown in Fig. 2, and the second set calculated by GainRatio shown Fig. 3.

D. The best reduced dataset

Now there are three datasets:

- First, Dataset#1, which is based on features permissions ranked by InfoGain, as shown in Fig. 2.
- Second, Dataset#2, which is based on features permissions ranked by GainRatio, as shown in Fig. 3.
- Third, the original Dataset based on all permissions.

To select the best reduced features set, WEKA [32] tool is used to evaluate the two sets against different classifiers.

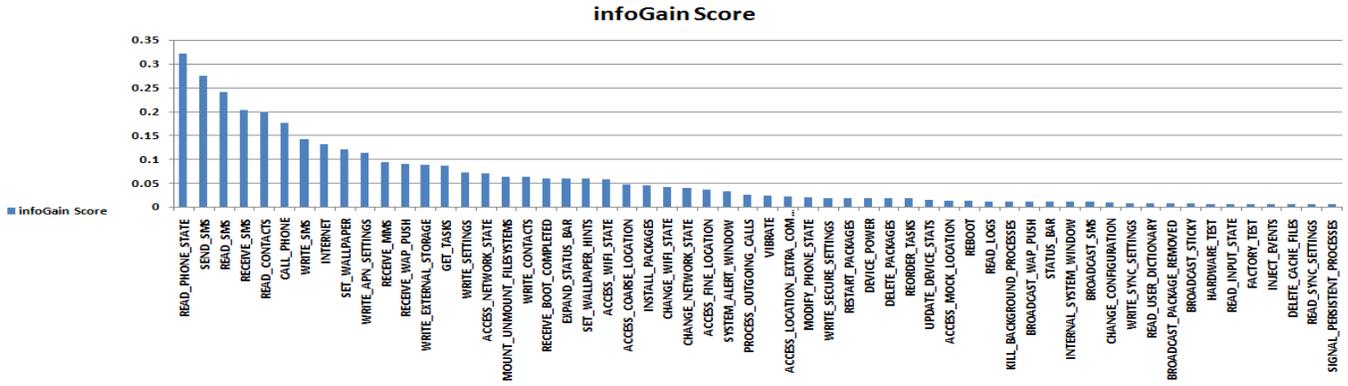


Fig. 2 Top 58 ranked feature using InfoGain

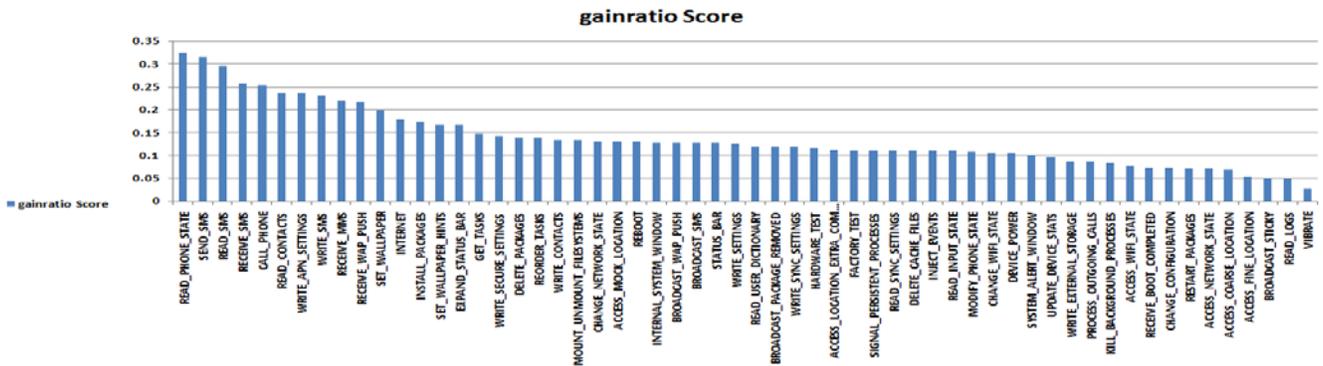


Fig. 3 Top 58 ranked feature using GainRatio

VI. EXPERIMENTS

WEKA tool is used to evaluate the three data sets against different classifiers, the results were compared that was obtained from all experiments from dataset based on all permissions and Dataset#1, which is based on features permissions ranked by InfoGain Dataset#2, which is based on features permissions ranked by GainRatio

A. Classifiers

More than one classifiers from WEKA are used to evaluate the best features set. The used classifiers are: C4.5 algorithm (J48), feed forward artificial neural network (MultilayerPerceptron MLP), Support vector machine (LibSVM), Radial Basis Function Network (RBFClassifier), stochastic gradient descent (SGD), Logistic Regression (Logistic), ExtraTree, J48Consolidated, RandomForest Tree, RandomTree, K-nearest neighbours classifier (IBk), KStar and best-first decision tree (BFTree)

B. Testing Options

WEKA has different mechanisms to divide the experimental dataset into training dataset and testing dataset testing that is

used to train and test the classifiers models. The first methodology is k-cross validation [33]. In k-fold cross-validation, the dataset is randomly partitioned into k equal size subsamples. One subsample is used as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. Then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged to produce a single estimation. In the conducted experiments, two k values have been chosen, k=10 and k=3 folds.

Another mechanism is simply to divide dataset into two portions in a random manner. Here, 66% of the original dataset are randomly chosen for training, and the remaining 34% of the data are used for testing and the last testing option is the use of training data as the testing data.

C. Measure of classifiers

The evaluation was performed by measuring the following

$$\text{metric: Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (5)$$

The *Accuracy* is the percentage of predictions that is correct, where TN is the number of benign applications correctly classified, TP is the number of malware cases

correctly classified, FP is the number of benign applications incorrectly detected as malware, and FN is the number of malware incorrectly classified as benign applications (false negatives).

D. Experiments steps

First, the classifiers was trained using Dataset#1 that is based on reduced features selected by InfoGain, see Fig. 4.

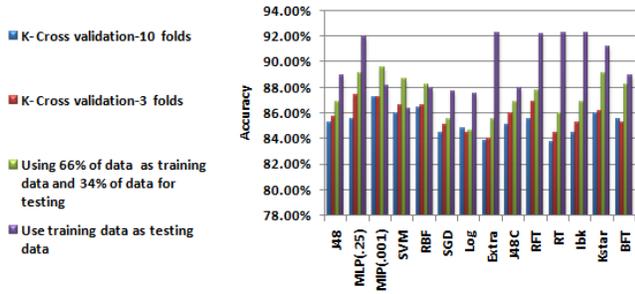


Fig. 4 the Accuracy for classifiers trained by Dataset#1 selected by InfoGain

Then, the classifiers was trained using Dataset#2 that is based on reduced features selected by gain ratio, see Fig. 5.

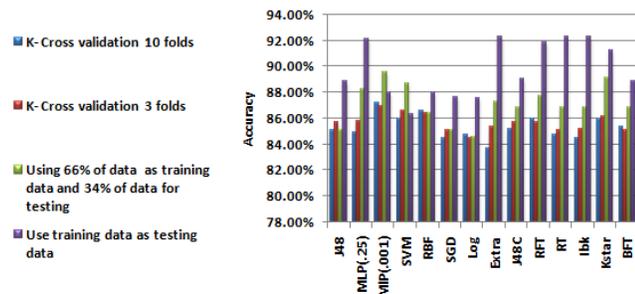


Fig. 5 the Accuracy for classifiers trained by Dataset#2 selected by GainRatio

Finally, the classifier was trained using original Dataset with all permission, see Fig. 6.

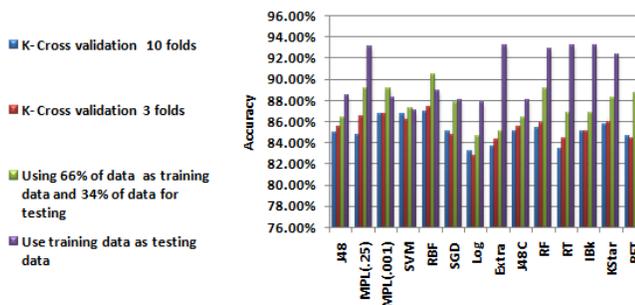


Fig. 6 The Accuracy for classifiers trained by Dataset with all permissions

VII. RESULTS

In this section, some of results obtained from previous experiments are concluded.

Fig. 7 show the results of the experiments conducted using k = 10 fold. The Fig. 7 show that in most of cases the classifications using the two reduced datasets give results better than that the classification using the dataset#3 gives. For example, with classifier MPL with learning rate .001 the reduced datasets give better result than using all permission, the classifier MPL (.001) with infoGain and RatioGain give accuracy 87.2308 % and with all permission give 86.7692 % and the best result given by all permissions given with RBF classifier 86.9231 %

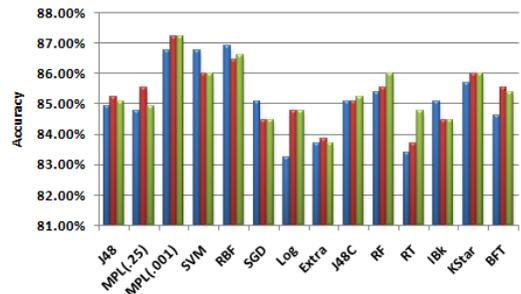


Fig. 7 Accuracy of classifiers using all permission and reduced permissions by InfoGain and GainRatio 10 folds

Fig. 8 show the results of the experiments conducted using k = 3 fold. Also, It is widely noted that classifications using the reduced datasets give results better that that given by classification using dataset#3 in most cases expect with SVM and RBF whereas the best result given with infoGain is the result given by MPL with leaning rate 0.25 is 87.3846 %, the best result given by GainRatio is given by MPL with learning rate .001 is 86.9231 % and the best rest result given by using all permission is given by RBF is 87.3846 %.

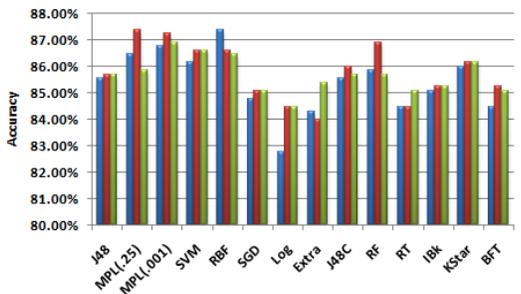


Fig. 8 Accuracy of classifiers using all permission and reduced permissions by info and GainRatio

But with using 34% as the testing option it is noted that in some classifier using all permissions give better result than reduced permissions and with other classifier the reduced permission give similar or better result than all permission see Fig. 9 whereas the best result given with infoGain and Gainratio is the result given by MPL with leaning rate 0.001 is

89.5928 %, and the best rest result given by using all permission is given by RBF is 90.4977 %.

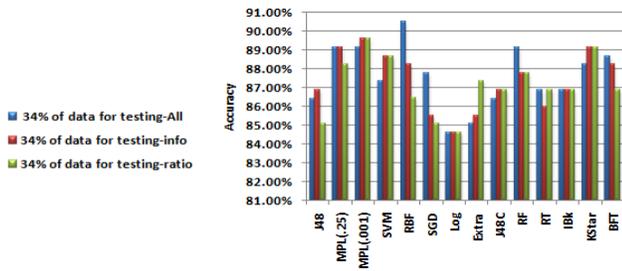


Fig. 9 Accuracy of classifiers using all permission and reduced permissions by info and ratio gain

So we can note from previous results in most cases especially with testing options 10 and 3 folds when we use the reduced permissions give equivalent or better results than using all permissions and with 34% as testing option the results are in with most classifiers are similar, so the reduced features set obtained by InfoGain or GainRatio can be used instead of using all android permissions as features for distinguish between benign and malware application.

The comparison between results obtained by info gain and gain ratio show that the accuracy for classifier trained by the feature set selected by info gain and gainratio are very similar as shown in Fig. 10 .

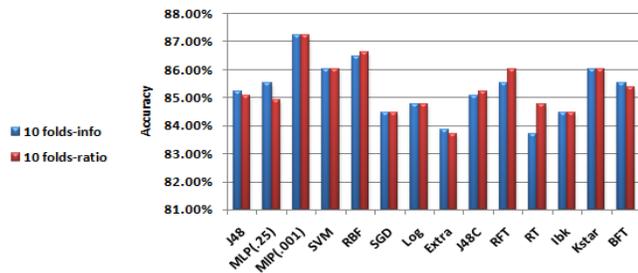


Fig. 10 InfoGain results against GainRatio result using 10 folds

But with the 3 fold and 34% testing options the result show the InfoGain result is better than the ratio results Except GainRatio give better results than InfoGain with RT and Extra Trees classifiers as shown in Fig. 11 and Fig. 12.

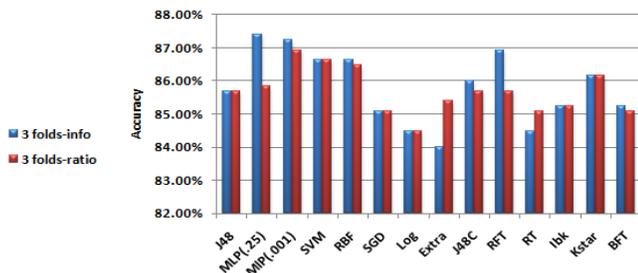


Fig. 11 InfoGain results against GainRatio result using 3 folds

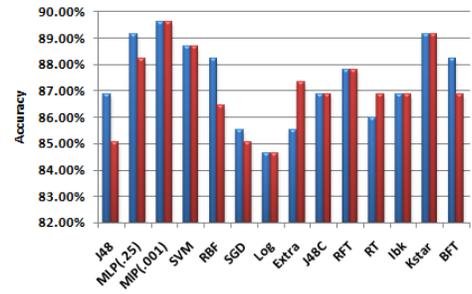


Fig. 12 InfoGain results against GainRatio result using 34% of dataset as a testing set

VIII. CONCLUSION

In this paper we used a android system permissions as features that extracted from Android application (.apk) files. The extracted data is used as features during a classification process of the applications. We focused on reducing number of features by selecting the best permissions that can distinguish between malware and benign applications.

We collected 325 android malware application and 325 benign applications and extracted the permissions and we reduced the number of permissions to 58 instead of using 151 permissions and select permissions based on InfoGain and GainRatio and test the two different set against different classifiers.

We concluded that the reduced permissions obtained by InfoGain or GainRatio , that extracted statically from .apk files, coupled with Machine Learning classifier can provide good indication about the nature of an .apk file without running it on the smartphone and it can be used instead of using all android permissions as features for machine learning classifier to distinguish between benign and malware application where the best result given by reduced permissions whether InfoGain or GainRatio is 87.2308 % and the best result obtained by using all permissions is 86.9231% .In the future work we will extract more features from applications to be combined with the reduced permissions .

ACKNOWLEDGMENT

We would to thanks everyone help us to complete this research and i would to thanks my supervisors Prof. Aliaa and Dr. Marwa for their excellent guidance.

REFERENCES

- [1] Smartphone Users Worldwide Will Total 1.75 Billion in 2014 [Online]. Available: <http://www.emarketer.com/Article/Smartphone-Users-Worldwide-Will-Total-175-Billion-2014/1010536>
- [2] Mobile Application Futures 2013-2017 [Online]. Available: <http://www.portioresearch.com/en/mobile-industry-reports/mobile-industry-research-reports/mobile-applications-futures-2013-2017.aspx>
- [3] Abdelfattah Amamra, Chamseddine Talhi, and Jean-Marc Robert, "Smartphone malware detection: From a survey towards taxonomy". In Malicious and Unwanted Software (MALWARE), 2012 7th International Conference on, pages 79–86. IEEE, 2012.

- [4] Mariantonietta La Polla, Fabio Martinelli, and Daniele Sgandurra, "A survey on security for mobile devices" *Communications Surveys & Tutorials*, IEEE, 15(1):446–471, 2013.
- [5] Suarez-Tangil, Guillermo, et al. "Evolution, detection and analysis of malware for smart devices." *Communications Surveys & Tutorials*, IEEE 16.2 (2014): 961-987
- [6] Yan Ma and Mehrdad Sepehri Sharbaf. "Investigation of static and dynamic android anti-virus strategies". In *Information Technology: New Generations (ITNG)*, 2013 Tenth International Conference on, pages 403–398 IEEE, 2013
- [7] Worldwide market share forecast of smartphone operating system from 2010 to 2015 [Online]. Available: <http://www.statista.com/statistics/266970/market-share-forecast-of-smartphone-operating-systems-from-2010-to-2015/>
- [8] Global Smartphone unit shipments forecast by operating system 2014 and 2018 [Online]. Available : <http://www.statista.com/statistics/309448/global-smartphone-shipments-forecast-operating-system/>
- [9] Cisco: 2014 Cisco Annual Security Report [Online]. Available: <http://www.cisco.com/web/offers/lp/2014-annual-security-report/index.html>
- [10] Android's Google Play beats App Store with over 1 million apps, now officially largest [Online]. Available: http://www.phonearena.com/news/Androids-Google-Play-beats-App-Store-with-over-1-million-apps-now-officially-largest_id45680.
- [11] Burguera, Iker, Urko Zurutuza, and Simin Nadjm Tehrani. "Crowdroid: behavior-based malware detection system for android." *Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices*. ACM, 2011.
- [12] Dini, Gianluca, et al. "Madam: a multi-level anomaly detector for android malware." *Computer Network Security*. Springer Berlin Heidelberg, 2012. 240-253.
- [13] Yerima, Suleiman Y., Sakir Sezer, Gavin McWilliams, Igor Muttik. "A new android malware detection approach using bayesian classification." *Advanced Information Networking and Applications (AINA), 2013 IEEE 27th International Conference on*. IEEE, 2013.
- [14] Sanz, Borja, et al. "Puma: Permission usage to detect malware in android." *International Joint Conference CISIS'12-ICEUTE'12-SOCO'12 Special Sessions*. Springer Berlin Heidelberg, 2013.
- [15] Liu, Xing, and Jiqiang Liu. "A Two-Layered Permission-Based Android Malware Detection Scheme." *Mobile Cloud Computing, Services, and Engineering (MobileCloud)*, 2014 2nd IEEE International Conference on. IEEE, 2014.
- [16] Android Operating System [Online]. Available : <https://www.android.com/>
- [17] Android Security [Online]. Available : <https://source.android.com/devices/tech/security/>
- [18] Android SDK [Online]. Available : <http://developer.android.com/sdk/index.html>
- [19] Android system Permissions [Online]. Available : <http://developer.android.com/guide/topics/security/permissions.html>
- [20] Contagio Mobile Mini Malware Dumb [Online]. Available : <http://contagiominidump.blogspot.com/>
- [21] Android Malware Dump [Online]. Available: <http://androidmalwaredump.blogspot.com/>
- [22] MalShare [Online]. Available : <http://malshare.com/>
- [23] Cooper, Vanessa N., Hossain Shahrar, and Hisham M. Haddad. "A Survey of Android Malware Characteristics and Mitigation Techniques." *Information Technology: New Generations (ITNG)*, 2014 11th International Conference on. IEEE, 2014.
- [24] Le Thanh, Hieu. "Analysis of Malware Families on Android Mobiles: Detection Characteristics Recognizable by Ordinary Phone Users and How to Fix It." *Journal of Information Security* 4.04 (2013): 213.
- [25] Google Play store [Online]. Available : <https://play.google.com/store>
- [26] Python 2.7 [Online]. Available : <https://www.python.org/download/releases/2.7/>
- [27] Androguard Project [Online]. Available : <https://code.google.com/p/androguard/>
- [28] Android Lollipop 5 [Online]. Available : <http://www.android.com/versions/lollipop-5-0/>
- [29] List of Android Manifest Permissions [Online]. Available : <http://developer.android.com/reference/android/Manifest.permission.htm>
- [30] Karegowda, Asha Gowda, A. S. Manjunath, and M. A. Jayaram. "Comparative study of attribute selection using gain ratio and correlation based feature selection." *International Journal of Information Technology and Knowledge Management* 2.2 (2010): 271-277.
- [31] T. M. Cover, J. A. Thomas, *Elements of Information Theory*, Ed. Wiley, 1991.
- [32] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1.
- [33] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." *Ijcai*. Vol. 14. No. 2. 1995

Ahmed Hesham Mostafa is from Cairo, Egypt. Ahmed was born in 23-Aug-1990, Ahmed is a 2011 graduate from faculty of computer science, Helwan University with degree in computer science with Excellent and honor degree. Ahmed completed the pre-master program in computer science in 2014.

He completed the military service in 2013, He is a Teacher Assistant in computer science department, Helwan University. He assisted in teaching many computer science programs such as Data structures programming, Logic Design, Computer architecture, Algorithms, Software Engineering, Design pattern in Java and Assembly Programming.

Ahmed is interested in Machine Learning, Data Mining, Malware and Mobile and computer security.

Marwa Mohamed Abd El Fatah, Assistant Professor of computer science, Faculty of computers and information, Helwan University, Cairo, Egypt.

She received her B.Sc and M.Sc. Degree in computer science from Helwan University. Dr. Marwa received the PHD degree in computer science from Helwan University in 2012. Field of interest includes pattern recognition, AI researches and mobile security.

Aliaa A. A. Youssif, Professor of computer science and vice dean for postgraduates and researches at Faculty of computers and information, Helwan University, Cairo, Egypt.

She received her B.Sc and M.Sc. Degree in telecommunication and electronics engineering from Helwan University. Prof. Aliaa received the PHD degree in computer science from Helwan university in 2000. She was a visiting professor at George Washington University (Washington DC, USA) in 2005. Field of interest includes pattern recognition, AI researches and medical imaging.

Fetal Heart Rate Estimation from Phonocardiograms Using an EMD Based Method

Dragos Daniel Taralunga, *Member, IEEE*, Mihaela Ungureanu, *Member, IEEE*, Bogdan Hurezeanu, *Member, IEEE*, and Rodica Strungaru, *Fellow, IEEE*

Abstract—Fetal monitoring is essential for evaluating the health status of the fetus during pregnancy. Long term fetal screening is recommended to reach a precise diagnostic. Acoustic noninvasive measurements of the fetal heart sounds by placing sensors on the maternal abdomen represent an attractive alternative to the standard cardiocography procedure used nowadays in hospitals, but the main problem is that these acoustic measurements are heavily corrupted by different types of noises, advanced signal processing methods being necessary for estimation of the fetal heart rate (fHR) from fetal phonocardiograms (fPCG). In this paper, a new method for fHR extraction is proposed having the following blocks: an Empirical Mode Decomposition (EMD) based block used to suppress the noise contaminating the fPCG; a block that enhances of the first heart sound (S1), a logic block that estimates the position of S1 based on rules concerning the fetal heart and a block which estimate the fHR. The proposed algorithm is validated on simulated fPCG, showing a very high accuracy in locating the S1 and estimating the fHR.

Index Terms—fPCG, EMD, fHR

I. INTRODUCTION

Fetal heart rate (fHR) represents one of the most important parameter used nowadays in hospitals for clinical assessment of fetal-well being during pregnancy. The useful information extracted from the analysis of the fHR is derived mainly from the characteristics of the baseline, variability, accelerations and decelerations of the fHR patterns [1], [2], [3], [4].

The standard method for acquiring the fHR is by using Doppler ultrasound cardiocography which involves the placement of an ultrasound sensor on the maternal belly. Ultrasounds are sent throw the maternal abdomen towards the fetal heart. However, this method has some drawbacks: it needs complex algorithms for fHR estimation (sometimes is likely that maternal heart rate (mHR) is reported as fHR) [5]; it does not offer information about the real operation of the valves, because it measures just the movement of the surface of fetal heart [6]; while long-term surveillance of the fHR is recommended, the impact of long prenatal ultrasound exposure is not yet clarified [7], [8], [9];

An alternative acoustic method is the use of fetal phonocardiography for fHR estimation. An acoustic (pressure) transducer is placed on the maternal abdomen which allows ones

This work was supported by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/132397

D. D. Taralunga, M. Ungureanu, B. Hurezeanu and R. Strungaru are with the Department of Applied Electronics and Information Engineering, Politehnica University Bucharest, Romania, e-mail: dragos.taralunga@upb.ro, bogdan.hurezeanu@gmail.com, mickyungureanu@yahoo.de, rodica.strungaru@upb.ro

Manuscript received June 22, 2015

to record the signal generated by the heart movement, i.e. the fetal phonocardiogram (fPCG). The advantages of this fetal monitoring procedure are: i) it is low cost, ii) no energy is transmitted to the fetus, the method being thus suitable for long-term surveillance, and iii) it handy to use, i.e. by placing a small acoustic sensor on the maternal abdomen, with no gel. Unfortunately, the low amplitude fetal cardiac signals measured on the maternal abdomen are heavily contaminated by different types of noise which can impair the proper analysis of the signal, consequently the fHR estimation. The main affecting factors that disturb the fPCG are: fetal movement, sensor movement, maternal heart sounds, respiratory sounds, maternal digestive sounds, acoustic attenuation due to amniotic liquid and other maternal tissues, external noise.

The early work in fetal phonocardiography addressed more the development of acquisition equipment and acoustic sensors [10], [11], [12], [13]. Chourasia *et al.* developed a wireless acquisition system using Bluetooth [14]. Recently, methods for fPCG processing are proposed in literature. Most of these methods usually apply band pass filters or Wavelet transform to reduce the noise in fPCG [15], [16], [17], [18], [19].

The current study proposes a new method for fHR estimation. The denoising is realized based only on the local characteristics of the contaminated fPCG signal. Thus no external reference like template or wavelet mother function is needed. The locations of the first fetal cardiac sound (S1), which is determined by the closure of the mitral and tricupsid valves, are determined using a logic block based on rules concerning the fetal heart physiology.

The rest of the paper is organized into the following sections: Section 2 describes the dataset used and the proposed fPCG denoising algorithm, which enhances the S1 allowing better fHR estimation. The results are discussed in Section 3 and concluding remarks are provided in the last section.

II. METHOD

A. Data description

The fPCG data used in this study are generated based on the algorithm proposed by Cesarelli *et al.* [14]. Thus, the fPCG signal can be defined by:

$$y(t) = s(t) - n(t) \quad (1)$$

where $y(t)$ is the composite signal including of the fetal heart sounds, $s(t)$ and the different types of noise, $n(t)$. The first and second sounds, S1 and S2, respectively, are modeled by two

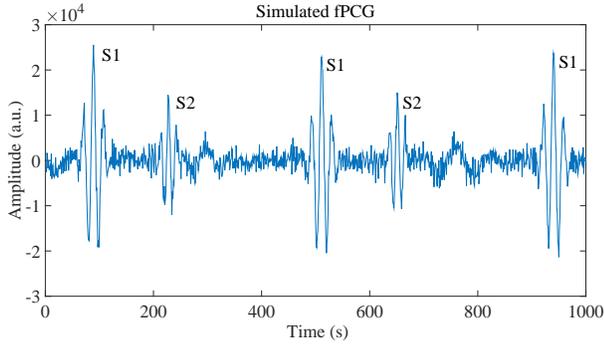


Fig. 1. Simulated fPCG signal with noise

Gaussian-modulated sinusoidal pulses [20]. The inter-distance between S1 and S2 is given by the following expression, [21]:

$$SSID = 210 - 0.5 * fHR \quad (2)$$

A detailed discussion about the simulation algorithm and about the simulated noise sources is presented in [22]; the simulated fPCG signals are available in [23]. In Fig. 1 a simulated fPCG signal with noise is depicted.

B. Proposed algorithm

The proposed algorithm is described by the blocks depicted in Fig. 2.

1) *Preprocessing and denoising*: Because S1 has higher energy as compared to the other fPCG components and less morphologic variability, it is appropriate for heart beat identification [24]. In order to reduce the background noise the Empirical Mode Decomposition (EMD) is used [25]. Thus, the fPCG is decomposed into a collection of intrinsic mode functions (IMFs). The IMFs containing noise are discarded and the remaining ones are used to reconstruct the signal.

The IMFs computation is fully data driven and the EMD algorithm has the following steps:

- 1) find all the local extrema of the data;
- 2) construct the upper and lower envelopes;
- 3) find the zero crossing;
- 4) calculate the mean of the upper and lower envelopes, $m_1(t)$, and subtract it from the original time series $x(t)$:

$$p_1(t) = x(t) - m_1(t) \quad (3)$$

- a) if $p_1(t)$ satisfies the conditions of the IMF, then $p_1(t)$ is the first IMF.
- b) if $p_1(t)$ does not satisfy the conditions of the IMF,

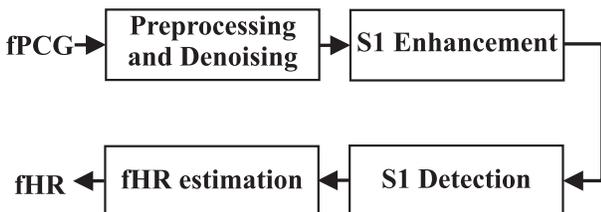


Fig. 2. Block diagram of the proposed algorithm

then steps 1), 2) and 3) are repeated for $p_1(t)$ and $p_{1,1}(t)$ is obtained:

$$p_{1,1}(t) = p_1(t) - m_{1,1}(t) \quad (4)$$

The above steps are repeated and after k cycles the IMF is obtained:

$$p_{1,k}(t) = p_{1,k-1}(t) - m_{1,k}(t) \quad (5)$$

Thus, the first frequency and amplitude modulated oscillatory mode, from the original data is obtained, $c_1(t) = p_{1,k}(t)$;

- 5) once the IMF is identified it is subtracted from the original data, obtaining the residual component $r_1(t)$:

$$r_1(t) = x(t) - c_1(t) \quad (6)$$

- 6) in the next cycle $r_1(t)$ is considered to be the original data. The above steps are repeated for l times, where l is the number of IMF plus the residue $r_l(t)$. A stopping criteria of the sifting process is defined in (7) and was introduced in [25] and is based on the standard deviation computed for two consecutive sifting results:

$$SD = \sum_{t=0}^T \frac{|p_{1,k-1}(t) - p_{1,k}(t)|^2}{p_{1,k-1}(t)^2} \quad (7)$$

The sifting process stops when the SD is within a predefined range, usually 0.2 - 0.3 [25].

The data can be reconstructed from its IMFs as following:

$$x(t) = \sum_{i=1}^l c_i(t) + r_l(t) \quad (8)$$

The first five IMFs are discarded because the high frequency noise is separated in the first IMFs. The auxiliary signal obtained after the reconstruction contains only the S1 and S2 and is further used in the next block for S1 enhancement.

2) *S1 enhancement and its detection*: The enhancement of S1 involves the following two steps: firstly the signal is squared and secondly, in order to obtain an envelope, a FIR filter with a Hamming window of 20 samples is applied and the delay introduced by this step is corrected. The resulted signal consists of very clearly delimited spikes corresponding to S1. A threshold is used to find possible candidates for S1:

$$T = \frac{0.7}{N} \sqrt{\sum_{i=1}^N b(i)^2} \quad (9)$$

where N is the total number of samples, and b is the signal resulted after the S1 enhancement block.

3) *fHR estimation*: In order to find the real S1 from the candidates obtained in the previous step, the following rules are applied:

- 1) if the time between two S1 candidates is smaller than 0.2 s, then the one with the lowest amplitude is rejected;
- 2) if the time between two S1 candidates is higher than 0.6 s, then a S1 was missed. The location of this missed S1 is estimated by adding to the position of current S1 the mean value of the difference between the last eight S1 correctly detected in the previous step.

The fHR is estimated using the following expression:

$$fHR = \frac{60}{T_{b,b}} (bpm); \quad (10)$$

where $T_{b,b}$ is the time between two consecutive S1. An fHR trace is generated by recording the fetal beat-to-beat values of fHR.

III. RESULTS

The method was applied on a 60 s simulated fPCG signal. In Fig. 3 is depicted the simulated fPCG signal, with noise, and in Fig. 4, is illustrated the signal after the denoising block based on the EMD is illustrated.

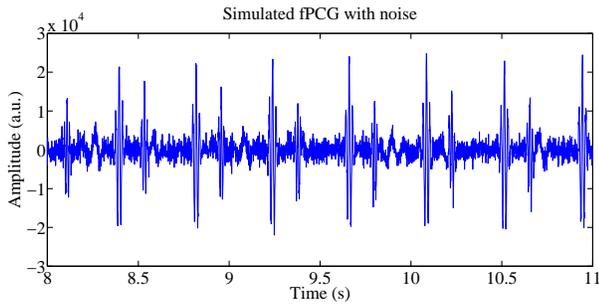


Fig. 3. Simulated fPCG signal with noise

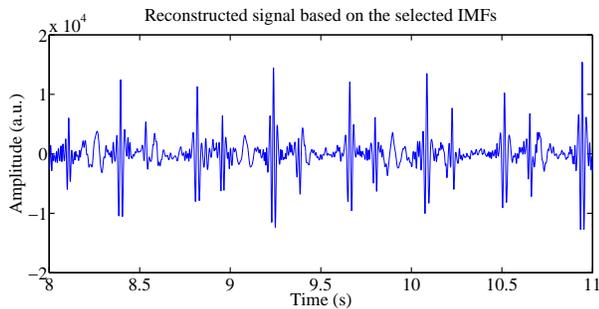


Fig. 4. Denoised fPCG signal using the EMD

The signal obtained after S1 enhancing, i.e. after squaring and smoothing, is depicted in Fig. 5.

In Fig. 6 a) and b) the robust detection of the S1, identifying its envelope and the corresponding peaks are shown.

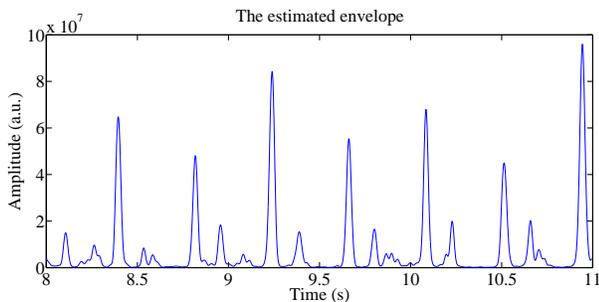


Fig. 5. S1 enhancement, detection of the envelope

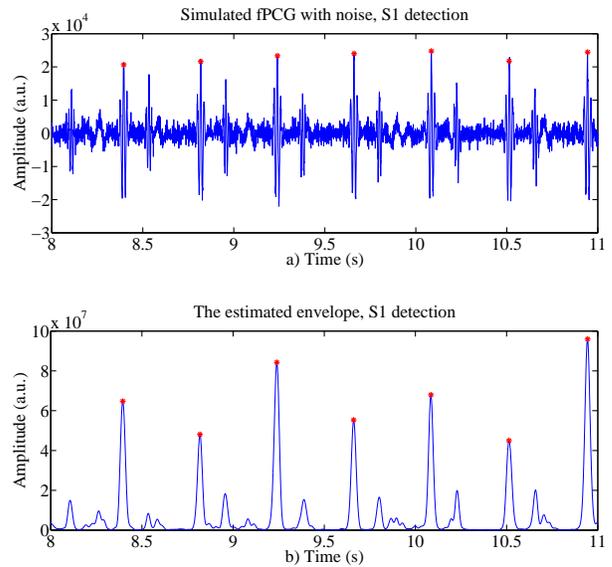


Fig. 6. a) S1 detection in the original fPCG; b) S1 detection in the envelope

Fig. 7 depicts the estimated fHR. The black arrow indicates the outliers which can appear when a S1 is omitted. It is recovered with the second rule from the fHR estimation block and its location is considered to be the location of the current S1 plus the mean S1 interval for the last eight detected S1.

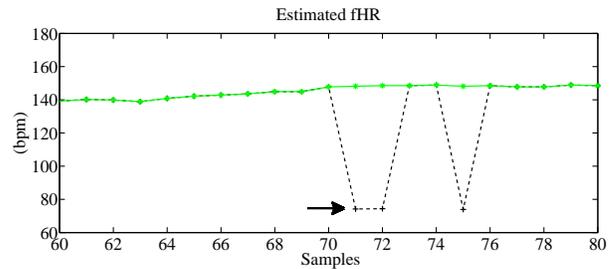


Fig. 7. fHR estimated from fPCG signal before the second rule of the fHR estimation block is applied. The arrow shows extreme values of fHR detected as outliers. These samples were substituted by samples depicted with *.

IV. CONCLUSION

The fPCG, generated by valves opening and closing, that allow the blood flow through the fetal heart, represents a passive, noninvasive recording method which permits the fHR estimation, being an attractive alternative to CTG. In order to obtain a good fHR estimate, a reliable detection of the fPCG characteristic points is mandatory. This can raise difficult signal processing issues due to the very low fPCG SNR and due to the morphology of the heart sounds. Moreover, the respiration affects the morphology of the heart sounds. However, mostly the S2 is affected, thus S1, having also a higher amplitude, is appropriate for fHR estimation. In this paper, a new method for S1 detection is presented, which considers the EMD as a preprocessing and denoising step. The method is able to detect all S1 even from noisy data, being very robust. Thus, reliable fHR is obtained, which can be used for further analysis. Future research will concentrated on making

the method fully automated, i.e., the automate selection of the IMFs to be discarded will be implemented. Moreover, the method will be validated on real fPCG signals.

ACKNOWLEDGMENT

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/132397.

REFERENCES

- [1] M. Fox, S. Kilpatrick, T. King, and J. T. Parer, "Fetal heart rate monitoring: Interpretation and collaborative management," *The Journal of Midwifery & Womens Health*, vol. 45, no. 6, pp. 498–507, 2000.
- [2] L. D. H. J. A. S. HB Krebs, RE Petres, "Intrapartum fetal heart rate monitoring. i. classification and prognosis of fetal heart rate patterns," *American Journal of Obstetrics and Gynecology*, vol. 133, no. 7, pp. 762–772, 1979.
- [3] A. Heazell, "Peripartum and intrapartum assessment of the fetus," *Anaesthesia & Intensive Care Medicine*, vol. 14, no. 8, pp. 333 – 336, 2013, regional Anaesthesia.
- [4] J. Parer, T. Ikeda, and T. King, "The 2008 national institute of child health and human development report on fetal heart rate monitoring," *Obstetrics and Gynecology*, vol. 114, no. 1, pp. 136–1368, 2009.
- [5] W. R. Cohen, S. Ommani, S. Hassan, F. G. Mirza, M. Solomon, R. Brown, B. S. Schiffrin, J. M. Himsworth, and B. R. Hayes-Gill, "Accuracy and reliability of fetal heart rate monitoring using maternal abdominal surface electrodes," *Acta Obstetrica et Gynecologica Scandinavica*, vol. 91, no. 11, pp. 1306–1313, 2012.
- [6] F. Kovacs, C. Horvath, A. T. Balogh, and G. Hosszu, "Fetal phonocardiographypast and future possibilities," *Computer Methods and Programs in Biomedicine*, vol. 104, no. 1, pp. 19 – 25, 2011.
- [7] H. Shankar, M.B.B.S. and P. P. Pagel, M.D., "Potential adverse ultrasound-related biological effectsa critical review," *The Journal of the American Society of Anesthesiologists*, vol. 115, no. 5, pp. 1109–1124, 2011.
- [8] E. S. B. C. Ang, V. Gluncic, A. Duque, M. E. Schafer, and P. Rakic, "Prenatal exposure to ultrasound waves impacts neuronal migration in mice," vol. 103, no. 34, pp. 12903–12910, 2006.
- [9] H. Kieler, S. Cnattingius, B. Haglund, J. Palmgren, and O. Axelsson, "Ultrasound and adverse effects," *Ultrasound in Obstetrics and Gynecology*, vol. 20, no. 1, pp. 102–103, 2002.
- [10] B. Tan and M. Moghavvemi, "Real time analysis of fetal phonocardiography," in *TENCON 2000. Proceedings*, vol. 2, 2000, pp. 135–140 vol.2.
- [11] P. Varady and L. Wildt, "Fetal phonocardiography with a novel approach," *Orvosi hetilap*, vol. 142, no. 36, pp. 1971–1976, 2001, cited By 3.
- [12] A. K. Mitra and N. K. Choudhari, "Development of a low cost fetal heart sound monitoring system for home care application," *Biomedical Science and Engineering*, vol. 2, no. 6, pp. 380–389, 2009.
- [13] F. Kovacs, M. Torok, C. Horvath, A. T. Balogh, T. Zsedrovits, A. Nagy, and G. Hossza, "A new, phonocardiography-based telemetric fetal home monitoring system," *Telemedicine and e-Health*, vol. 16, no. 8, pp. 878–882, Oct. 2010.
- [14] V. S. Chourasia and A. K. Tiwari, "Wireless data acquisition for fetal phonographic signals using bluetooth," *International Journal of Computers in Healthcare*, vol. 1, no. 3, pp. 240 – 253, 2012.
- [15] V. S. Chourasia and A. K. Mitra, "A comparative analysis of denoising algorithms for fetal phonocardiographic signals," *IETE Journal of Research*, vol. 55, no. 1, pp. 10–15, 2009.
- [16] ———, "Selection of mother wavelet and denoising algorithm for analysis of foetal phonocardiographic signals," *Journal of Medical Engineering & Technology*, vol. 33, no. 6, pp. 442–448, 2009.
- [17] F. Kovacs, C. Horvath, A. Balogh, and G. Hosszu, "Extended non-invasive fetal monitoring by detailed analysis of data measured with phonocardiography," *Biomedical Engineering, IEEE Transactions on*, vol. 58, no. 1, pp. 64–70, Jan 2011.
- [18] V. S. Chourasia and A. K. Tiwari, "Design methodology of a new wavelet basis function for fetal phonocardiographic signals," *The Scientific World Journal*, vol. 2013, p. 12, 2013.
- [19] V. S. Chourasia, A. K. Tiwari, and R. Gangopadhyay, "A novel approach for phonocardiographic signals processing to make possible fetal heart rate evaluations," *Digital Signal Processing*, vol. 30, no. 0, pp. 165 – 183, 2014.
- [20] A. Mitra, N. Choudhary, and A. Zadgaonkar, "Development of an artificial womb for acoustical simulation of mothers abdomen," *International Journal of Biomedical Engineering and Technology*, vol. 1, no. 3, pp. 315–328, Jan. 2008.
- [21] F. Kovacs, M. Torok, and I. Habermajer, "A rule-based phonocardiographic method for long-term fetal heart rate monitoring," *Biomedical Engineering, IEEE Transactions on*, vol. 47, no. 1, pp. 124–130, Jan 2000.
- [22] M. Cesarelli, M. Ruffo, M. Romano, and P. Bifulco, "Simulation of foetal phonocardiographic recordings for testing of {FHR} extraction algorithms," *Computer Methods and Programs in Biomedicine*, vol. 107, no. 3, pp. 513 – 523, 2012.
- [23] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [24] P. Varady, L. Wildt, Z. Benyo, and A. Hein, "An advanced method in fetal phonocardiography," *Computer Methods and Programs in Biomedicine*, vol. 71, no. 3, pp. 283 – 296, 2003.
- [25] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

Soft-error-rate Adaptive Intervals for Low Overhead Checkpoint

Wentao Jia, Chunyuan Zhang and Kun Jiang

Abstract—Soft errors are an increasingly important threat to the reliability of integrated circuits. Chips manufactured in advanced technologies show variations in SER caused by variations in the process parameters. Ongoing reduction of feature sizes and complexity of operating environment (temperature, voltage, radiation pressure and so on), SER variation is increasingly manifesting.

Checkpoint is one of the most popular recovery method used for many systems, and the intervals of checkpoint can obviously influence performance. However, optimal intervals of checkpoint rely on SER. Theoretically speaking, SER adaptive checkpoint(SACP) which dynamically match checkpoint intervals with real time SER can improve checkpoint overhead under variable SER. But benefits of SACP are relative with SER variation.

We give a mathematical model of SER variation and proposal a way to predict SER based errors occurred most currently. Results show high accuracy of SER prediction and much overhead improvement of SACP.

Keywords—soft-error-rate, adaptive, error variation, checkpoint intervals, low overhead

I. INTRODUCTION

Soft errors including SBU(single-bit-upset) and MBU(multiple-bit upset) are an increasingly important threat to the reliability of integrated circuits fabricated in advanced CMOS technologies. Researchers expect an aggregate effect on soft-error rate (SER) of a chip. The error rate at 16-nm may be almost 100 times that at 180 nm [1].

Ongoing reduction of feature sizes has increased the probability that a single particle causes an MCU(Multiple-cell upset). It has been observed that a single neutron caused more than 50 bit flips in an SRAM of 65nm [2]. During neutron-accelerated SER tests of SRAMs in advanced processes, it is not unusual to observe that more than 50% of the upset events are MCUs [2], implying ECC techniques generally applied are more and more not capable of protecting systems.

Chips manufactured in advanced technologies gradually show variability on SER caused by variation in the process parameters. Voltage supply variability, voltage threshold variability and channel length variability at the lower level are

directly associated with overall circuits at the higher level, not only performance variability and power variability, but also SER variability. Process variability causes that the SER vulnerability of an SRAM bit cell is not the same for its two data states. Experimental results for a 90-nm embedded SRAM showed that the differences can be almost a factor of 4 [3]. SRAM bit cells are symmetric by design, however sequential logic are not. As a result, the SER of sequential logic usually varies with the data state(0/1) and clock state(HIGH/LOW), the differences can be nearly 10X [4].

SER differences caused by process variability while chips manufactured and designed will eventually show SER variations while systems running, meaning SER of systems may not keep constant but change from time to time. Moreover, SER is highly associated with operating environment (temperature, voltage, radiation pressure and so on), which is not constant always but variable sometime. As computing is emerging anywhere and any-time, operating environment of many systems are variable, such as spacecraft, smart city system, unmanned verticraft, intelligent vehicle. As cars drive from New York City to Denver, CO, USA, SER increase 3.5X (due to altitude increased by 1.6KM) [5]. Many systems employ dynamic mechanisms to decrease power or energy, like DVFS, which also making voltage or frequency variable. Ongoing reduction of feature sizes and complexity of operating environment, SER variation is increasingly manifesting.

The most popular recovery method is checkpoint, and intervals of checkpoint can obviously influence performance. Based on checkpoint model proposed by Daly [6], Figure 1 shows how the performance overhead varies with checkpoint intervals(section II-B shows more details of Daly's model). Optimal checkpoint interval for MTTE(mean time to errors)=100min is 20min. Keeping the interval while SER changing to MTTE =10min or MTTE=1000min, overhead will be 450% or 11%. However, overhead at optimal interval is 115% or 7%. So SER adaptive intervals of checkpoint can reduce overhead for variable SER. Theoretically speaking, SER adaptive checkpoint(SACP) which dynamically match checkpoint intervals with real time SER can improve checkpoint overhead under variable SER.

There are two points to reduce overhead for present checkpoint mechanism: reducing MTTE or time overhead to update checkpoint. Differently, the view of SER adaptive checkpoint (SACP) is that analyze occurrence of errors more carefully and match checkpoint interval with real time SER dynamically. Benefit of SACP is relative with SER variation (how large the variation takes and how long it keeps), so we firstly have to evaluate impact of SER variability on SER adaptive checkpoint. We make the following contribution in this paper:

W. Jia is with National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, ChangSha, 410073 China (e-mail:wjia@nudt.edu.cn).

C. Zhang is with National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, ChangSha, 410073 China (e-mail:cyzhang@nudt.edu.cn)

K. Jiang is with National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, ChangSha, 410073 China (e-mail:kunjiang@nudt.edu.cn).

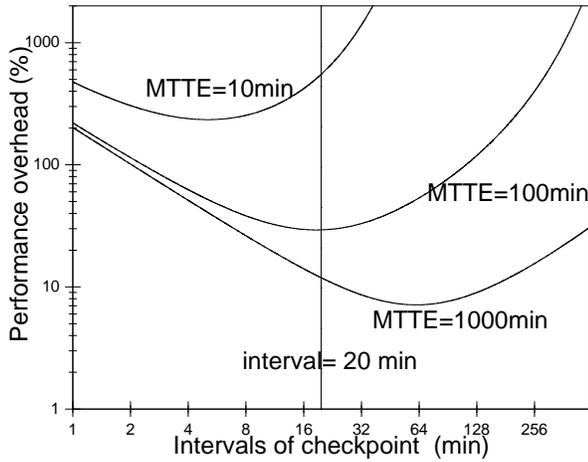


Fig. 1. checkpoint overhead

- Based on how SER affected by temperature, voltage, radiation pressure and so on, we give a mathematical model with four parameters to quantized SER variation.
- We study impact of variable SER on overhead of present checkpoint and SER adaptive checkpoint, indicating optimal benefits of SACP.
- We proposal a way to predict SER based dynamic prediction window, showing practical benefits of SACP.

Section II gives background information on SER variation and related works on checkpoint. Section III models SER variation. Section IV introduces our way to predict SER. Section V provides details on evaluation methods and results are given in Section VI.

II. BACKGROUND AND RELATED WORKS

This paper focus on decreasing checkpoint overhead under variable SER. SER can be described by mean time to errors(MTTE) or number of errors occurred in unit time, and this paper will use later, that is $SER=1/MTTE$.

A. SER variation.

Soft errors are caused by either neutrons generated by cosmic radiation interacting with the earth's atmosphere or alpha particles emitted by radioactive impurities that are present in chips and package materials. SER can be get from the following equation [7]:

$$SER = Constant \times Flux \times Area \times e^{Q_{crit}/Q_{coll}}$$

Flux is the alpha or neutron flux experienced by the circuit, Area is the effective diffusion area, Q_{crit} is Critical charge, and Q_{coll} is the collection efficiency. Flux varies with altitude, location on the earth, concrete thickness of building, solar activity and so on. Q_{crit} depends on the supply voltage and temperature. Table I lists variable factors related to SER.

Form table I, we can see SER exponentially varies with most variable factor, for variable factor $X \in$

TABLE I. VARIABLE FACTORS RELATED TO SER

variable factor X	how SER varies with X	amplitude of SER variation ^a
Altitude	$SER0 \times e^{0.9x-0.03x^2}$ [5]	over 100(sea level to flight level 12KM)
Latitude		3 (20N-60N)
Concrete Thickness	$SER0 \times e^{-2.7x}$ [8]	15 (0-1m)
Solar Cycle		1.3[8](day to next day) 1.6[8](month to next month)
Voltage	$SER0 \times e^{-1.3x+0.3}$ [9]	3 (0.4V-1.0V)@45nm
Temperature	$\approx exp$ [10]	3^b (25°C -110°C)

^adefined as maximal SER divided by minimum SER

^boperational temperatures of chips are distributed in a wide range and much higher than the ambient temperature in reality

{Altitude, Thickness, Voltage, Temperature}, SER could be

$$SER(X) = SER0 * e^{C_x*(X-X_0)} \quad (1)$$

X_0 and SER_0 are a pair known value of X and corresponding SER, C_x is a constant indicating amplitude of SER variation with X variation per unit. Take altitude for example, $C_x=0.8$, meaning SER will increase $e^{0.8}$ per kilo-meter higher. If $C_x < 0$, indicating a decrease of SER while X increase, such as voltage and thickness. In summary, amplitude of SER variation caused by operating environment can be from several to a hundred and caused by process variability can be ten. Considering both, amplitude of SER variation can be nearly thousand.

B. Checkpoint

Checkpoint technique is commonly used to recover from application failure. During checkpointing shown in Fig.2, application entire statement is written to storage so that in occurrence of errors application can resume its work from the last checkpoint rather than from the beginning. T_{ex} in Fig.2 is execution time of application, T_{ov} is time to store application statement, T_r is time to recovery application, T_{solve} is solve time for application which is equal to $N*T_{ex}$, N is the number of passed segments required to complete a calculation. One aspect of employing checkpoint is properly assigning checkpoint intervals T_{ex} . Daly[6] proposed a accurate ways based

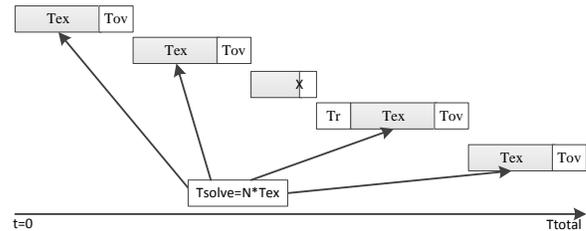


Fig. 2. The application time line broken into four passed compute segments and one failed compute segment designated by X. An application run is complete when the accumulated computation time T_{ex} of all of the passed segments is equal to the total solution time T_{solve} for the application.

SER and T_{ov} to determine the optimum checkpoint interval.

$$T_{ex} = \begin{cases} \sqrt{\frac{2T_{ov}}{SER}} - T_{ov} & \text{for } T_{ov} \leq 1/2SER \\ 1/SER & \text{for } T_{ov} > 1/2SER \end{cases} \quad (2)$$

Solve time in checkpoint is time spent on actual computational cycles, so checkpoint overhead is

$$overhead = \frac{T_{total} - T_{solve}}{T_{solve}} \times 100\%$$

We also use Daly's method to determine checkpoint intervals in this paper.

C. Adaptive checkpoint

Only a few works focus on adaptive checkpoint, for some argue that checkpoint interval has a little impact on applications. Jones[12] studied impact of checkpoint intervals with relative error from 0.1 to 4 on application efficiency. As his results, efficiency decrease from 58% to 50% while relative error of checkpoint intervals is 100%, demonstrating that checkpoint intervals has little impact on application efficiency. However, we focus on overhead, which will increase by 28% while efficiency decreases from 58% to 50%. Moreover, relative error of checkpoint intervals caused by SER variation may be much more than 10.

Previous research improving overhead focused primarily on decreasing the checkpoint time of data transferred, while relying on constant checkpoint frequency. Gerofi [13] proposed a algorithm that adapts dynamically to the properties of the workload being executed, such as changes in the number of dirtied memory pages, network and disk I/O operations, as well as to the network bandwidth available for replication. The results show benefits of adaptive checkpoint.

The most difference with previous works is that we are focusing on SER variation which has rarely been studied.

III. MODELLING SER VARIATION

We define SER at time t as SER(t), let initial SER be SER0, so constant SER is

$$SERc(t) = SER(0) = SER0$$

Variable SER according to Eq. (1) is

$$SERv(t) = SER0 * e^{Cx*(X(t)-X0)} \quad (3)$$

X(t) usually relevant to system type, application field, operating environment and so on. Variation of X(t) based X0 is shown as Eq.(4)

$$X(t) = \begin{cases} X0, & \text{for } (t\%T) \leq T - Tv \\ X0 + a \times \sin(\frac{\pi(t-T+Tv)}{Tv}), & \text{for } (t\%T) > T - Tv \end{cases} \quad (4)$$

with a indicating the largest amplitude increased or decreased, T is cycle time of variation and we use the mod % operator to make it, Tv is time for variation amplitude a (shown in EQ.4) sustained. Here we use sin() function to describe variation for a complicated variation can be regarded as some simultaneous simple sin() variation. Figure 3 gives the X(t) as a=-0.5 and a=1. Take X(t) to Eq.(3),we have

$$SERv(t) = \begin{cases} SER0, & \text{for } (t\%T) \leq T - Tv \\ SER0e^{Ax\sin(\frac{\pi(t-T+Tv)}{Tv})}, & \text{for } (t\%T) > T - Tv \end{cases} \quad (5)$$

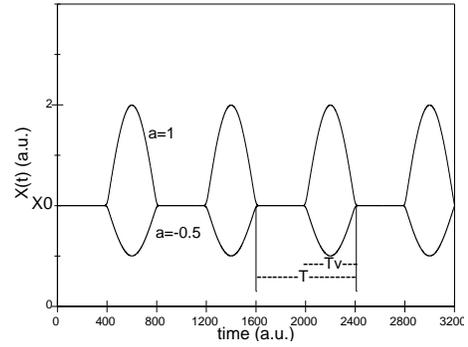


Fig. 3. examples of X(t) variation

Here we make A=Cx*a, Cx is SER variation per unit X(t) variation, a is X(t) total variation, so A is SER total variation for X(t). Take altitude as an example, Cx=0.8(SER increased with e^{0.8} time per kilo-metre increased), suppose a=5KM, so A=4, meaning SER increased to e⁴ time while X(t) increased. Figure 4 gives SERv(t) examples for A=-1, A=0.5 and A=3.

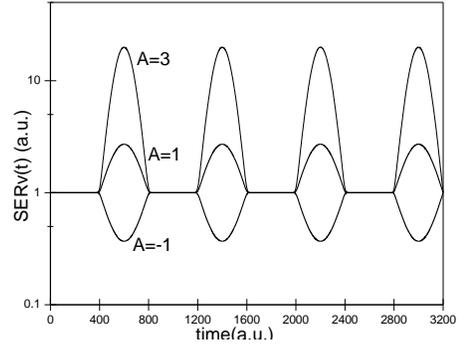


Fig. 4. examples of SER(t) variation

We can get number of errors occurred one cycle T by cumulative SERv(t) from 0 to T, and we can see number of errors at A=3 is greater than A=0, and that at A=0 is also greater than A=-1. Overhead is greatly relative with errors number, and this paper is going to study overhead due to SER variability, so we use a factor K to equalize errors of SERv(t) in Eq.(6).

$$SERv(t) = \begin{cases} SER0 * K, & \text{for } (t\%T) \leq T - Tv \\ SER0 * Ke^{Ax\sin(\frac{\pi(t-T+Tv)}{Tv})}, & \text{for } (t\%T) > T - Tv \end{cases} \quad (6)$$

Number of errors in every T for constant SER is $\int SERc(t)dt = SER0 * T$. we equalize errors for SERv(t) as following function

$$\int_0^T SERv(t)dt * K = SER0 * T$$

So,

$$K = \frac{SER0 * T}{\int SERv(t)dt}$$

and value of K is determined by A and T_v/T . Figure 5 shows the $SER_v(t)$ equalized.

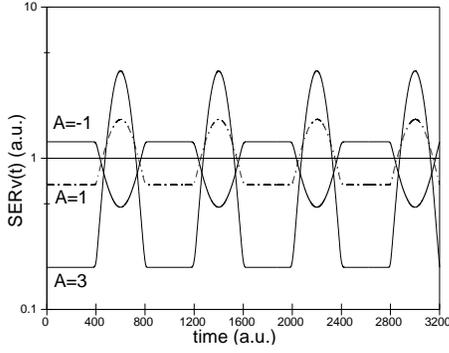


Fig. 5. examples of $SER(t)$ variation equalized

SER variation we studied in this paper is based on Eq.(6), we summary the parameters as following:

- SER_0 : the average SER in every cycle T .
- A : amplitude of SER variation, defined as maximal SER /minimum $SER=e^{|A|}$. We categorize SER variation as increased variation if $A > 0$ and decreased variation if $A < 0$.
- T_v : T_v is duration for variation sustained.
- $pn= T_v/T$ is time proportion of variation, T is cycle time of variation.

Our goal is to study how the parameters (SER_0, A, T_v, pn) impact checkpoint overhead for increased and decreased variation respectively. SER variation actually is more complicated than we modelled in Eq.(6), however, we can regard a complicated variation as some simultaneous simple variation.

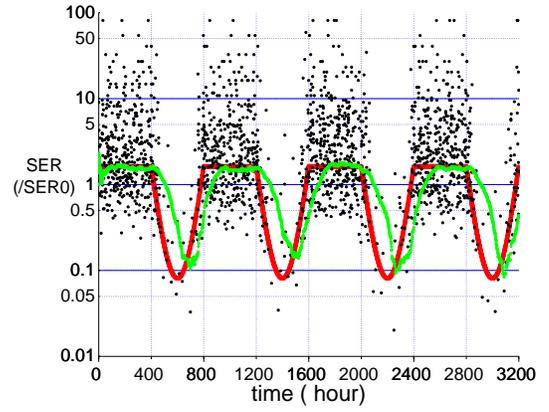
IV. SER PREDICTION

SER is usually calculated by number of errors occurred/time systems have run. Due to variation, we could not use errors occurred long time ago to predict real-time SER , just as Fig 6(a) shown. Simulated results in Fig 6(a) manifest delay between theoretical SER and predicted SER using errors occurred in 200 hours. So the key to SER prediction is proper time of errors occurred, and we define periods employed to predict SER as prediction window(PW).

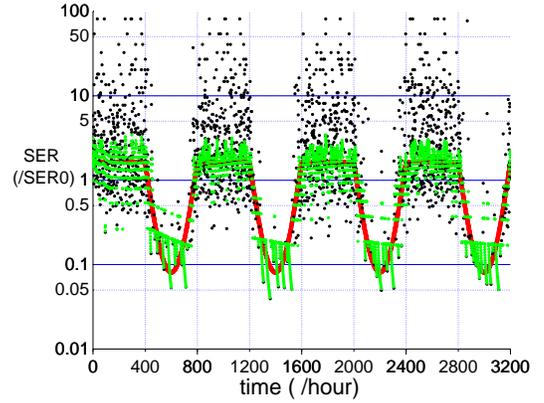
Two reasons decreasing accuracy of SER prediction is random of errors occurred and delay due to SER variation. Errors occurred randomly(as Fig.6(b)) and we need many errors to amortize deviation of SER . Whereas we should use errors occurred as current as possible for less delay deviation. Unfortunately, the two ways to decrease deviation is usually contrary. More errors to amortize random deviation indicate longer time ago errors to be used, which will obviously worsen prediction accuracy under low SER .

To balance the deviation cause by delay and random, we employ shorter PW when SER is largely changing and longer PW when SER relative constant, as shown in table II.

DPW is dynamically increase or decrease the prediction window based $RE(s_1, s_2)$ and $numbErr(pw)$, making it works



(a) prediction window=200h, showing deviation cause by delay



(b) prediction window=10h, showing deviation cause by random

Fig. 6. SER prediction. Red line is theoretical SER as Eq.(6), black points are simulated errors occurrence based red line, green points are SER we predicted based black points. Suppose $SER_0 = 1/100min^{-1}$.

well for both continuous and variation. DPW is low-overhead and need no more system information, except the time of each error occurred.

We use error ratio between predicted SER and theoretical SER (shown as Eq.(6)) to measure precision of prediction, Error ratio of at t is defined as

$$ER(t) = \frac{MAX(SER_{pre}(t), SER_{OPT}(t))}{MIN(SER_{pre}(t), SER_{OPT}(t))}$$

and we refer error ratio to overestimating if $SER_{pre}(t) > SER_{OPT}(t)$ or underestimating if $SER_{pre}(t) < SER_{OPT}(t)$. Sum of ER as entire application ran is

$$ER = \sum_{i=1}^{i=N} ER(t_i) \times \frac{T_{ex_i}}{T_{solve}} \quad (7)$$

here N is number of checkpoint intervals, t_i is the time i th interval started, $T_{ex}(i)$ is length of i th interval, T_{solve} is $\sum_{i=1}^{i=N} T_{ex_i}$. Sum of ER is great than 1, the larger means the worse precision of prediction.

TABLE II. ALGORITHM OF DYNAMIC PREDICTION WINDOW (DPW)

Function:	
numbErr(pw): number of errors occurred during prediction window	
SER(pw):SER calculated by numbErr(pw)/pw	
RE(s1,s2): relative error of s1 and s2, abs(s2-s1)/s1	
input:	
t: current system time	
Nt: number of errors occurred at current time	
Time[Nt]: time the ith error occurred, used for numbErr(pw)	
output: SER(t)	
1	MTTE \leftarrow t/Nt
2	pw \leftarrow 10 \times MTTE
3	WHILE numbErr(pw) > 50
4	pw \leftarrow pw - MTTE
5	IF pw < 10 \times MTTE
6	RETURN SER(t) \leftarrow SER(pw)
7	s1 \leftarrow SER(pw), s2 \leftarrow SER(pw + MTTE)
8	WHILE RE(s1, s2) < 1 AND (pw) < 50
9	pw \leftarrow pw + MTTE
10	RETURN SER(t) \leftarrow SER(pw)

V. SIMULATION METHODS

We simulate the checkpoint mechanism as following steps in table III and simulation code is listing in appendix A.

TABLE III. METHOD TO SIMULATE CHECKPOINT

Function:	
Tex: checkpoint intervals	
computTex(ser):calculate Tex based SER	
Parameter:	
Ttotal : total simulated time	
step: minimum time unit for simulation	
output: Tsolve.(overhead = (Ttotal - Tsolve)/Tsolve)	
1	WHILE timeSIM < Ttotal
2	timeSIM \leftarrow timeSIM + step
3	IF error occurred in this step
4	Rflag \leftarrow 1 // start to recovery
5	ELSE IF Rflag==1 and recovery finished
6	Tex \leftarrow computTex(ser)
7	Rflag \leftarrow 0
8	ELSE IF checkpoint interval is finished
9	Tsolve \leftarrow Tsolve + Tex
10	Tex \leftarrow computTex(ser)

Issues we need to determine is: how less should simulation step be(which will determine the simulation precision) and how many times should the simulation need to repeat for each case (which will average the simulation deviation). The simulation generates pseudo-random errors in a Poisson distribution. Based on [12], [6] basic parameters for checkpoint are given: $SER0 = 1/100min^{-1}$, $Ttotal = 3200h$ (so number of errors occurred in theory is $3200 * 60 / 100=1920$), time overhead to update checkpoint is 2min, time to recovery errors is 4min. T and Tv are relative to SER0 (enough errors should occur to show SER variation in every T), here we suppose $T=800h$, $Tv=400h$.

The step of simulation has great influence on result precision. For Poisson distribution simulated, only one error occurred per simulation step at most, so relative error of number occurrence is $(\lambda * step - 1 * (\lambda * step)e^{-\lambda * step}) / (\lambda * step) = 1 - e^{-\lambda * step}$. Given $\lambda = SER0 = 1/100min^{-1}$, when step=1min, 0.1min, 0.01min, relative error is respectively 0.99%, 0.10%, 0.01%, so we donate step=0.1min to expect relative error of number occurrence less than 0.1%. However, SER variation will affect the precision. Table IV shows that mean error number under step=0.1 min is fit well for

TABLE IV. ERRORS SIMULATED PRECISION FOR DIFFERENT AMPLITUDE A. MEAN NUMBER OF ERRORS SHOULD BE 1920

step(min)	mean errors number and relative error(%)			
	A=0	A=1	A=3	A=5
0.1	1918.73 0.07%	1917.74 0.12%	1911.71 0.43%	1906.56 0.70%
MIN{0.1, 0.1*SER0/SERv(t)}	1918.21 0.09%	1918.97 0.05%	1918.1 0.10%	1918.54 0.08%

A=0(relative error less 0.1%), but not good for A=5(relative error 0.7%), in which errors may occur much more in short time. We used a viable step for high SER, as step is minimum of 0.1 and $0.1 * SER0 / SERv(t)$. The viable step fit very well for all value of A, making relative error less than 0.1%.

How many times we should simulated per case to average simulation result? According to [11], we denote 99% confidence level, 0.1% relative error, and get times>3458. We simulated 3500 times per case, so we can assure a 0.1% precision for errors simulated.

We simulated respectively for the following methods:

- FIX, fixed interval of checkpoint, the intervals are computed with the SER0 and never change during system execution.
- SACP_OPT, SER adaptive intervals with the theoretical SER as Eq.(6), we use SACP_OPT to evaluate the optimal benefit of SACP.
- SACP_DPW, SER adaptive interval with SER predicted by DPW, in order to show the benefit can be achieved by SACP.

VI. RESULTS

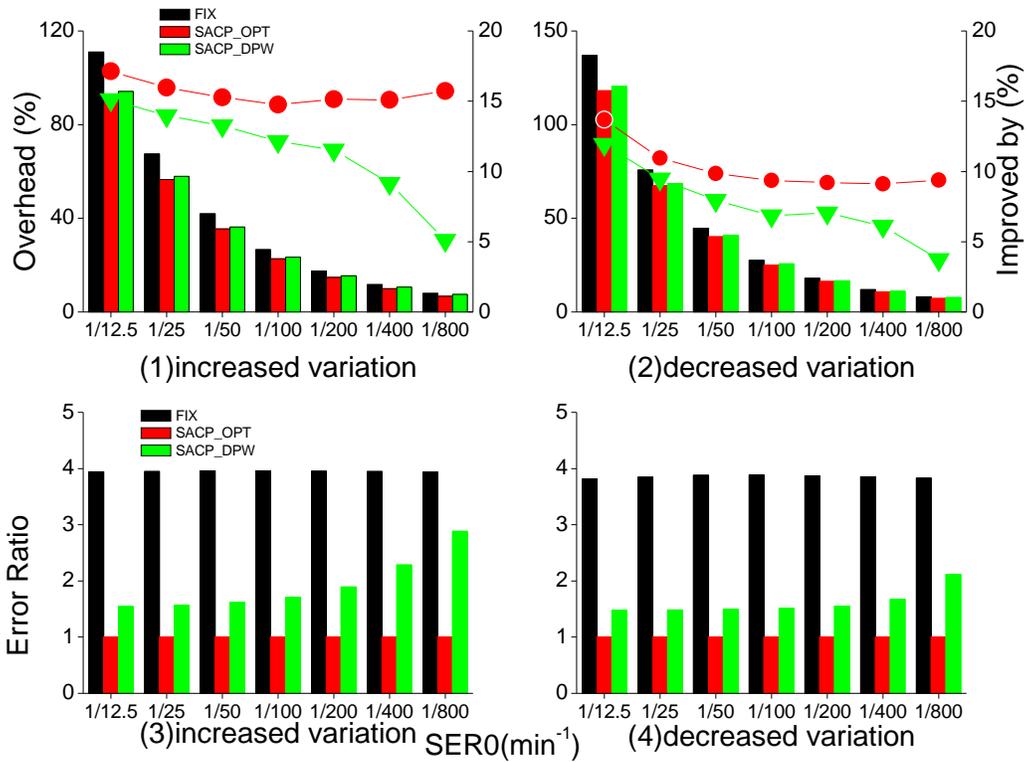
A. SER0

Figure 7(a) shows how overhead vary with SER0 from $1/800min^{-1}$ to $1/12.5min^{-1}$. Results in Fig. 7(a)(1) show that overhead benefit between SACP_OPT and FIX quickly increase from 1.3 % to 19%. Overhead improvement of SACP_OPT increase from 15% ($SER0 = 1/100min^{-1}$) to 17% ($SER0 = 1/12.5min^{-1}$). Results also show improvement nearly constant with SER less than $1/100min^{-1}$, that is to say SACP_OPT can always improve overhead by 15 % even for very low SER.

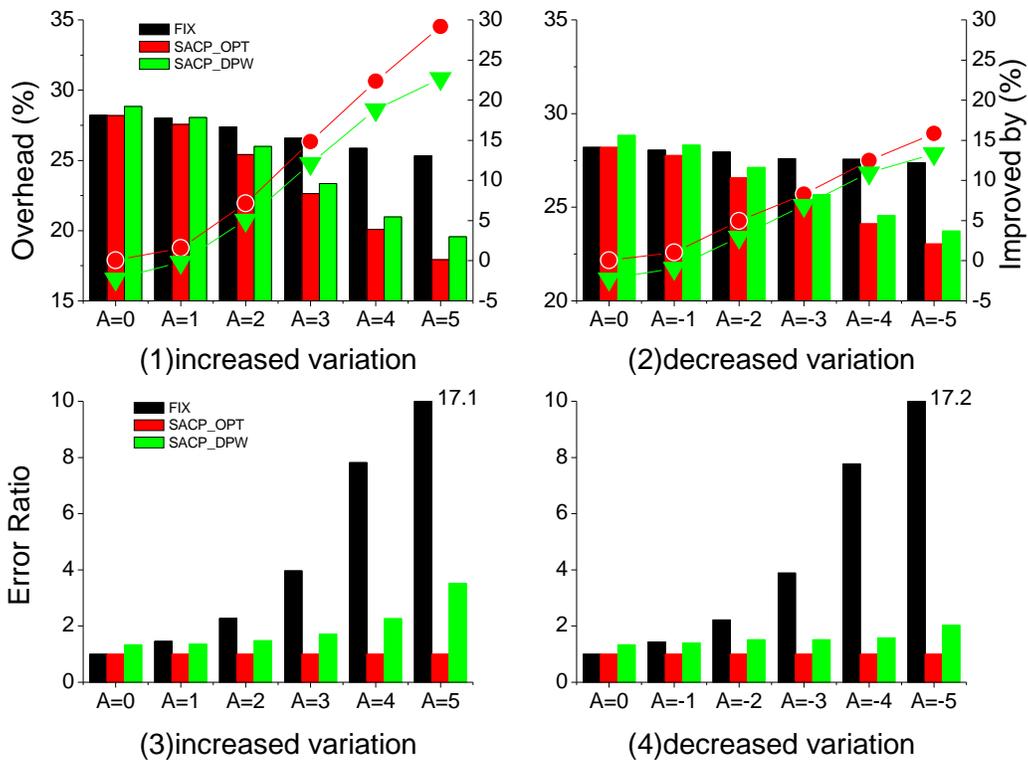
Overhead improvement of SACP_DPW largely increases form 5% to 15% as SER0 higher. SACP_DPW hardly improves overhead with SER less than $1/800min^{-1}$. because DPW predicts real-time SER based on errors, lower SER, less errors occurred at same periods, less precision for SER prediction.

We recall error ratio(ER) defined in Eq.(7) that ER of SACP_OPT is always 1 and more ER means more relative error of SER compared with SACP_OPT. Error ratio in Fig. 7(a)(3) clearly show that ER of SACP_DPW stays less than 2 with $SER0 < 200min^{-1}$ and increases to 2.8 in $1/800min^{-1}$ and probably is greater than FIX with $SER0 > 3200min^{-1}$ under which condition that so few errors occurred for SER prediction.

Fig. 7(a)(2)and(4) show the same thing, summarily increased variations get better overhead improvement than decrease but worse error ratio of SER prediction.

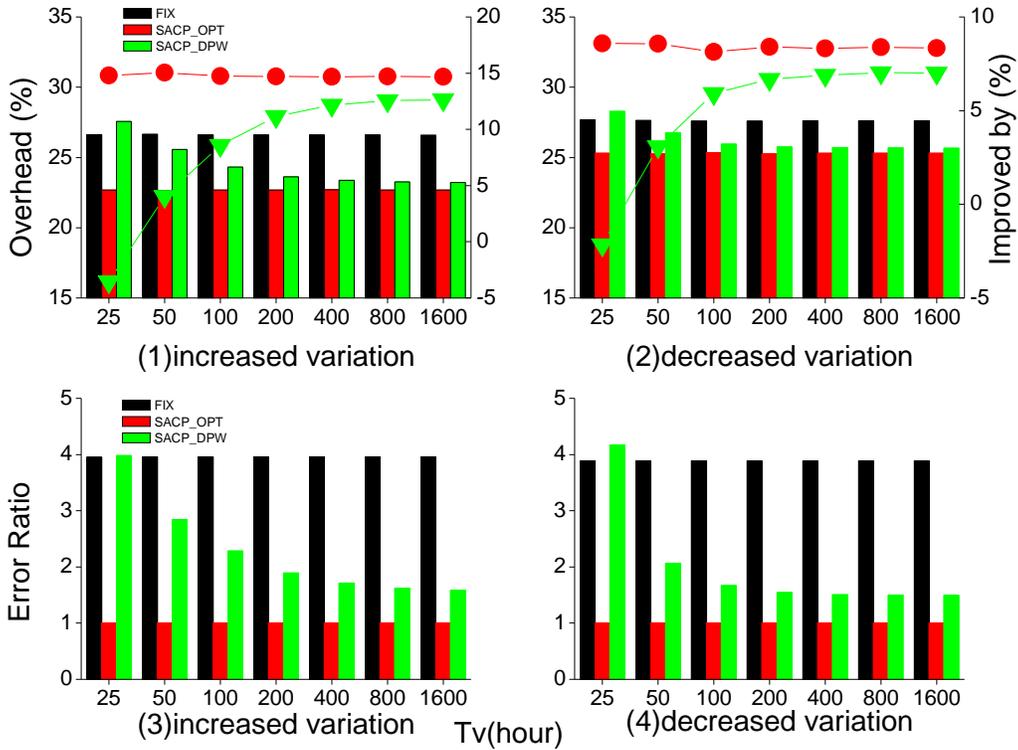


(a) overhead and corresponding improvement for $SER0$ with $|A| = 3, Tv = 400h, pn = 1/2$

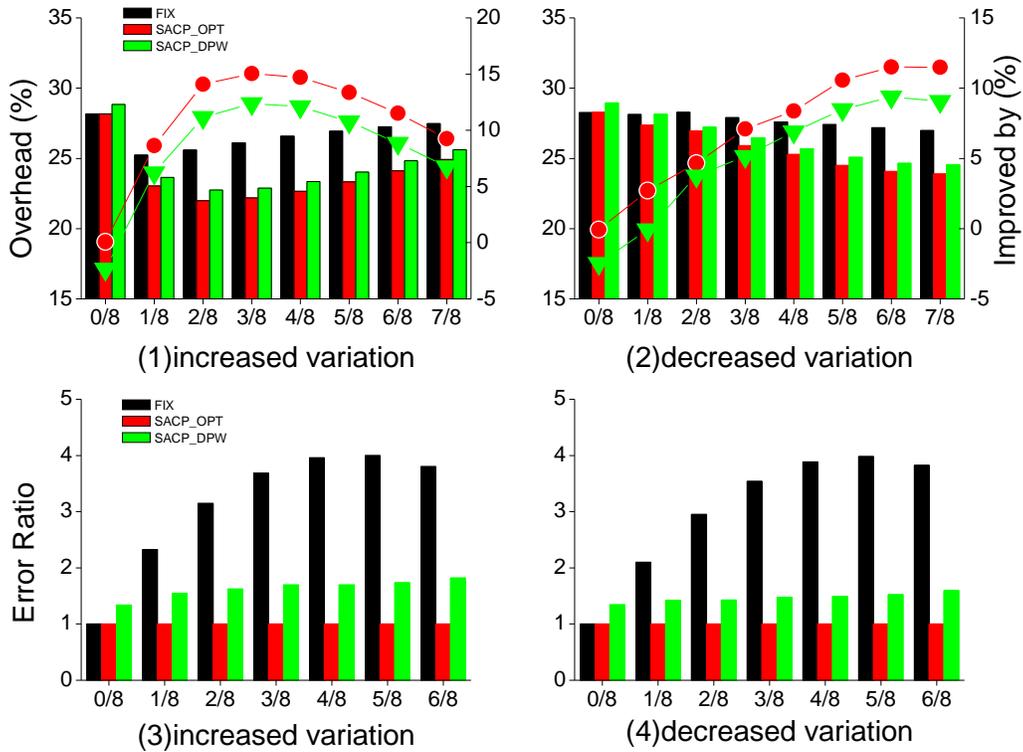


(b) overhead and corresponding improvement for A with $SER0 = 1/100\text{min}^{-1}, Tv = 400h, pn = 1/2$

Fig. 7. overhead(bar shown with left axis) and corresponding improvement (line shown with right axis) for (1)overhead and corresponding improvement for increased variation (2)overhead and corresponding improvement for decreased variation (3)error ratio for increased variation (4)error ratio for decreased variation.



(a) overhead and corresponding improvement for Tv with $|A| = 3, SER0 = 1/100min^{-1}, pn = 1/2$



(b) overhead and corresponding improvement for pn with $|A| = 3, SER0 = 1/100min^{-1}, Tv = 400h$

Fig. 8. overhead(bar shown with left axis) and corresponding improvement (line shown with right axis) for (1)overhead and corresponding improvement for increased variation (2)overhead and corresponding improvement for decreased variation (3)error ratio for increased variation (4)error ratio for decreased variation.

B. Variation amplitude A

Results in Fig. 7(b) show that as amplitude getting larger, overhead of SACP_OPT and SACP_DPW obviously decrease. When A less than 1, SACP_DPW is worse than FIX, while A larger than 2 SACP_DPW improve FIX overhead as much as by 25%. Results of error ratio show that DPW works well, specially for decreased variation with ER less than 2 for all A.

C. variation duration T_v

Results in Fig 8(a) show nearly constant improvement of SACP_OPT, while improvement of SACP_DPW obviously decrease as $T_v < 200h$. If T_v is less than 25h, SACP_DPW will get no improvement. So few errors occurred in such a short periods make the variation maybe finished before SACP_DPW senses the variation.

D. variation proportion p_n

Results in Fig 8(b) show benefits increase as p_n increase to 4/8, however, a more complicated trend after that for increased variation and decreased variation. The reasons are shown in Fig 9. Error ratio and overestimating error ratio

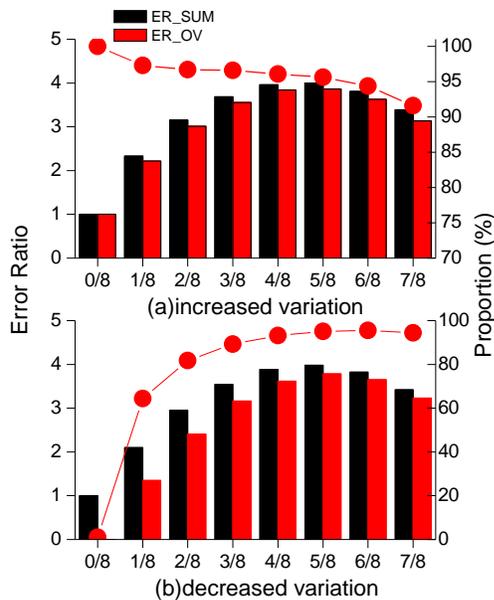


Fig. 9. error ratio of FIX for p_n . Error ratio (bar shown with left axis, sum of overestimating and underestimating error ratio labelled as ER_SUM and overestimating error ratio labelled as ER_OV) and overestimating proportion (line shown with right axis) for (a) increased variation (b) decreased variation

shown in black and red bar are increase for both variation, however, proportion of overestimating error ratio decreases from 100% to 92% for increased variation but increases from 0% to 95% for decreased variation. Just as many research mentioned overestimating checkpoint interval will result in more overhead than underestimating, that is why overhead of

increased variation and decreased variation show differently trend in Fig 8(b).

VII. CONCLUSION

Soft errors are an increasingly important threat to the reliability of integrated circuits. Chips manufactured in advanced technologies show variation in SER caused by variation in the process parameters. Ongoing reduction of feature sizes and complexity of operating environment, SER variation is increasingly manifesting. checkpoint is the most popular recovery method, and the intervals of checkpoint can obviously influence performance. But optimal intervals of checkpoint are determined on SER. SER adaptive checkpoint which dynamically match checkpoint intervals with real time SER can improve checkpoint overhead under variable SER. But benefit of SACP is relative with SER variation, so in this paper we evaluate impact of SER variability on SER adaptive checkpoint.

This paper is focusing on SER variation which has rarely been studied. Based on how SER affected by temperature, voltage, radiation pressure and so on, we model SER variation with four parameters: average SER SER_0 , variation amplitude A, variation duration T_v and variation proportion p_n . With a exact simulation, we come out how much each parameter impacts the performance overhead of present checkpoint and SER adaptive checkpoint. We proposal a way to predict SER based dynamic prediction window (DPW), showing practical benefits of SACP.

We categorize SER variation to increased variation and decreased variation. we give how much the parameters (SER_0, A, T_v, p_n) impact checkpoint overhead for those four variability respectively. Actual SER of systems may be more complicated than we modelled in this paper, however, we can regard a complicated variation as some simple variation acted simultaneously. Results in this paper can be referred to systems on weather SER variation need be considered or how much benefit SER adaptive checkpoint maybe achieved.

ACKNOWLEDGMENT

The authors gratefully acknowledge supports from National Nature Science Foundation of China under NSFC No. 61402504, 61033008, 61272145, and 61103080; National High Technology Research and Development Program of China under No. 2012AA012706.

REFERENCES

- [1] Borkar S. Designing reliable systems from unreliable components: the challenges of transistor variability and degradation. IEEE Micro, 2005, 25(16): 10-16
- [2] Nicolaidis M. Soft errors in modern electronic systems. New York, USA: Springer Science and Business Media, 2011.
- [3] Heijmen T. Spread in alpha-particle-induced soft-error rate of 90-nm embedded SRAMs. Proceedings of the 13th IEEE International On-Line Testing Symposium. Heraklion, Crete, Greece, 2007: 131136
- [4] Heijmen T, Roche P, Gasiot G, Forbes K R and Giot D. A comprehensive study on the soft-error rate of flip-flops from 90-nm production libraries. IEEE Transactions on Device and Materials Reliability, 2007, 7(1): 8496

- [5] Zielger J F and Puchner H. SER-history, trends, and challenges: a guide for designing with memory ICs. Cypress Semiconductor Corporation, 2004.
- [6] Daly J T. A higher order estimate of the optimum checkpoint interval for restart dumps. *Future Generation Computer Systems*, 2006, 22: 3033-12
- [7] Hazucha P and Svensson C. Impact of CMOS technology scaling on the atmospheric neutron soft error rate. *IEEE Transactions on Nuclear Science*, 2000, 47(6): 2586-2594
- [8] JEDEC standards on measurement and reporting of alpha particle and terrestrial cosmic ray induced soft errors in semiconductor devices. Arlington, VA: JEDEC Solid State Technology Association, JESD89, 2006
- [9] Chandra V, Aitken R. Impact of Technology and Voltage Scaling on the Soft Error Susceptibility in Nanoscale CMOS. *IEEE International Symposium on Defect and Fault Tolerance of VLSI Systems*, Boston, MA, USA, 2008: 114 - 122
- [10] Jagannathan S, Diggins Z, Mahatme N N, Loveless T D, Bhuvu B L, Wen S J, Wong R and Massengill L W. Temperature dependence of soft error rate in flip-flop designs. *IEEE International Reliability Physics Symposium*, Anaheim, CA, 2012: SE.2.1 - SE.2.6
- [11] Shubu Mukherjee. *Architecture design for soft errors*. San Francisco, CA: Morgan Kaufmann Publishers Inc, 2008.
- [12] Jones W M, Daly J T and Debardeleben N. Impact of sub-optimal checkpoint intervals on application efficiency in computational clusters. *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, New York, NY, USA, 2010: 276-279
- [13] Gerofi B, Ishikawa Y. workload adaptive checkpoint scheduling of virtual machine replication. *Proceedings of the 2011 IEEE 17th Pacific Rim International Symposium on Dependable Computing*, Washington, DC, USA , 2011: 204-213
- [14] Zhang Chengye, Deng Shenglan, Ning Hong. A local checkpoint mechanism for on-board computing. *Proceedings of 2012 IEEE International Conference on Information Science and Technology*, Hubei, China, 2012: 520-526.
- [15] Xiong Lei and Tan Qingping. Data flow error recovery with checkpointing and instruction-level fault tolerance. *2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies*, Gwangju, Korea, 2011: 79-85.

Wentao Jia was born in 1985. He is a Ph.D. candidate in School of Computer from National University of Defence Technology. His research interests include fault-tolerant computing and computer architecture.

Chunyuan Zhang was born in 1964. He is the professor in School of Computer from National University of Defence Technology. His research interests include computer architecture, parallel programming, low power design, embedded systems, media processing and scientific computing.

Kun Jiang was born in 1984. He is a Ph.D. candidate in School of Computer from National University of Defence Technology. His research interests include parallel programming and computer architecture.

A Novel Technique to Detect and Recognize Faces in Multi-view Videos

Steven Lawrence Fernandes, Dr. G. Josemin Bala

Abstract— In this paper we have developed two novel techniques to detect and recognize faces from Videos and across Obscure and Various Lighting Conditions. Detecting and recognizing face images from Videos is proposed using Spherical Harmonics and Radial Basis Function (RBF) kernel Technique and tested using three different videos. Recognizing images across Obscure and Various Lighting Conditions is proposed using Local Binary Pattern (LBP) Feature Extraction and Support Vector Machine (SVM) Classification Techniques and tested using Yale B database which consists of images across various lighting conditions. Both the proposed systems were tested across various existing state of the art techniques. From our analysis we have found that proposed detecting and recognizing faces from videos using Spherical Harmonics and RBF kernel Techniques gives the best recognition rate of 98% and Recognizing faces across Obscure and Various Lighting Conditions using LBP Feature Extraction and SVM Classification gives the best recognition rate of 100% on Yale B database.

Keywords— Face Detection, Face Recognition, Spherical Harmonics, Radial Basis Function Kernel, Local Binary Pattern, Support Vector Machine.

I. INTRODUCTION

Detection and Recognition of human faces across image processing are two difficult task mainly detection and recognition of human faces across videos and across blurring and illumination affects. Detecting and recognizing faces from video is hard because of its continuously changing views[1]-[2] and blurring and illumination makes it still more hard[3]-[4]. So we propose two novel techniques they are, Detecting and Recognizing Face Images from Videos and Recognizing Images across Obscure and Various Lighting Conditions. In case of Detection and Recognition of Face Images from Videos it is easy to identify individuals in frontal view cameras but it's difficult in other views, in other views it requires some tedious works including a multiple camera network, an effective tracking system to track and identify the persons in effective manner [5]-[7]. The identification of the persons in different views can be much helpful in identifying the suspicious persons in a video. Identification of human faces in video has numerous applications in video surveillance systems.

When it comes to Blur and Illumination face recognition alludes to the procedure of recognizing people in view of their facial features. In blurring the edge content of the image will

be reduced and the image colours will be smooth that it almost looks similar and recognition of parts or content of image will be difficult, whereas illumination variation that happens on face pictures debases the execution of face recognition system under pragmatic situations[8]-[14]. For face recognition variations across pictures of the same face due to illumination changes are bigger than image variations in pictures because of variations in identities of faces [15]-[19]. Chen et al. [20] delineated this unpredictability by demonstrating that there is no discriminative illumination invariant for Lambertian questions on source of light which are placed apart. Which means that it is impractical to figure out if two pictures were made by the same objects under two different sources of light or by separate objects? Therefore issued one picture of an object, it is hard to anticipate anything positive about this item or how will it show up under varying lighting conditions.

II. RELATED WORK

Here we have done some study of previous works on Detecting and Recognizing Images from Multi-Viewing Videos and Recognizing Images across Blurring and Illumination affects which was helpful in knowing more about these problems and various techniques to solve them. Some of the study contents are listed below.

A. Detecting and Recognizing Images from Multi-Viewing Videos

Authors in [1] developed the model that may be all around connected to remake 3D shape from pictures of a more extensive ethnic mixture. By converting the laser scans to vectors of texture and shape we can obtain a point to point of all the scans with respect to reference scan to ensure that same points are being described by vector dimensions like nose tip. Optical flow derives the algorithm which computes the correspondences automatically. Authors in [2] explained that a face detector is applied to the input image. For this two algorithms were employed, Support Vector Machine (SVM) and Boosting. The input image containing the best detection is cropped to a square region around the face, and is then rescaled. The second stage computes position estimates of eye and mouth corners (will be referred to as Component Of Interest (COI), hereafter) in the image. The algorithm predicts the position of a COI from pixelintensities within a $k \times k$ window. Invariance under intensity changes is achieved by subtracting the mean value from each window, and dividing it by its standard deviation. The Kernel Ridge Regression (KRR) is adopted for this purpose. To predict the position of that

component in a test image, all estimates are computed, and then binned into 1-pixel-sized bins. The bin with the most votes is chosen as the predicted location.

B. Recognizing Images across Blurring and Illumination affects

Authors in [3] developed a system where local descriptions are built by making use of texture descriptors and combining all these local descriptions to obtain global description. The facial image is divided into local regions and texture descriptors are extracted from each region independently. The Local Binary Pattern (LBP) labels contains information about the patterns on a pixel-level, we add these local information to obtain information in regional level and the regional level information's are concatenated to develop a global description of the face. Authors in [4] focused on local parts in face images. These images are taken under different pose, various lighting conditions and In the presence of some occlusions like sunglasses which may make it hard to identify an image with our naked eye.

III. PROPOSED SYSTEM

In this paper we have worked on two difficulties in recognizing faces through images. And have proposed two systems to work on these problems having one methodology for each problem. The proposed two methodologies are explained below.

A. Detecting and Recognizing Face Images from Videos using Spherical Harmonics and RBF kernel Techniques

The video is converted into frames. Frames are considered as images. Here we have two phases training phase and testing phase. Training phase consists of images from database which are not affected by any noise and whose face regions are defined clearly, whereas testing phase consists of images obtained from video which may have noise and whose face regions are not defined. Pre-processing is applied to the video frames to remove noise using median filter. Then face region in the frame is detected and is masked. Based on the position of the face in the frame the face region is detected throughout the video. Histogram of Oriented Gradients (HOG) features are extracted from images in both training and testing phase. Spherical Harmonics is applied to the images in the training and testing phase for recognition. The distance between the train image feature and the test image features were calculated. The image that is having the minimum distance is retrieved from the database.

Modules for Detecting and Recognizing Face Images from Videos using Spherical Harmonics and RBF kernel Techniques

- Pre-processing
- Face Detection
- Face Masking
- Tracking

- Recognition

Pre-processing

The video is first converted into frames. Each frame is considered as image. Noises in the images reduce the quality of the images. In order to improve the quality of the image Median filter is used for filtering. The median filter considers all pixels of an image. With one pixel as reference it observes the nearby pixel values arranges all the pixel values in ascending order and finds the median value. It then replaces the reference pixel value by this newly obtained median value. This process continues for all pixels.

Face Detection

The face is detected from the frame using the vision cascade operator which identifies the face region in the image. It gives the x and y position of the face images. The position of the detected face image is taken and a rectangle is drawn at the particular location.

Face Masking

The face region is then masked based on the detected face region. The region which is detected as face is taken and the particular region is marked separately. Areas other than the region which is detected as a face is blackened and face region alone is masked. This will be helpful in the correct recognition of the face images in later stages.

Tracking

The position of the face regions at each frame is updated each time and the person is tracked throughout the video. Every time the face region is identified by analyzing the movement of the person in the consecutive frame. The position is updated each time so that the system is trained to identify the movement of the person in the frame and all these times the position of the rectangle is moved according to the movements identified in the frame.

Recognition

Feature Extraction is done using HOG and these features are applied to Spherical Harmonics for recognition. The distance is calculated between the test feature and the train feature. The image corresponding to the feature having minimum distance is retrieved from the database.

- **Spherical Harmonics**

The spherical harmonics $Y_l^m(\Theta, \phi)$ are angular part of resolution to Laplace's equations in spherical coordinates.

Here, Θ is taken as the colatitudinal coordinate with $\Theta \in (0, \Pi)$, and ϕ as the longitudinal coordinate with $\phi \in (0, 2\Pi)$. Spherical harmonics satisfy the spherical harmonic differential equation, which is given by the angular part of Laplace's equation in spherical coordinates. Spherical Harmonics are defined as

$$Y_l^m(\theta, \phi) = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}} P_l^m(\cos \theta) e^{im\phi} \quad (1)$$

Where,

$$P_l^m(x) = (-1)^m (1-x^2)^{\frac{m}{2}} \frac{d^m}{dx^m} P_l(x) \quad (2)$$

Where,

$$P_l(x) = \frac{1}{2^l l!} \frac{d^l}{dx^l} (x^2 - 1)^l \quad (3)$$

Here m is order, l is degree

• **RBF kernel**

In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in support vector machine classification. RBF kernel on two samples x and x' is defined as

$$K(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \quad (4)$$

Where $\|x - x'\|^2$ a squared Euclidean distance and σ is a free parameter.

• **Bhattacharyya Distance**

The Bhattacharyya distance measures similarity of two discrete or continuous probability distributions. It is closely related to Bhattacharyya coefficient which is a measure of the amount of overlap between two statistical samples or populations.

Bhattacharyya coefficient of two distributions p and q in the same domain x is defined as,

$$BC(p, q) = \sum_{x \in X} \sqrt{p(x)q(x)} \quad (5)$$

And Bhattacharyya difference can be calculated as,

$$D_B(p, q) = \frac{1}{4} \ln\left(\frac{1}{4} \left(\frac{\sigma_p^2}{\sigma_q^2} + \frac{\sigma_q^2}{\sigma_p^2} + 2\right)\right) + \frac{1}{4} \left(\frac{\mu_p - \mu_q}{\sigma_p^2 + \sigma_q^2}\right)^2 \quad (6)$$

Where, $D_B(p, q)$ is Bhattacharyya distance between p and q classes

- σ_p is variance of pth classes
- μ_p is mean of pth class
- p and q are the two different distributors

System architecture of detecting and recognizing face images from videos using Spherical harmonics and RBF kernel techniques is shown in Fig.1.

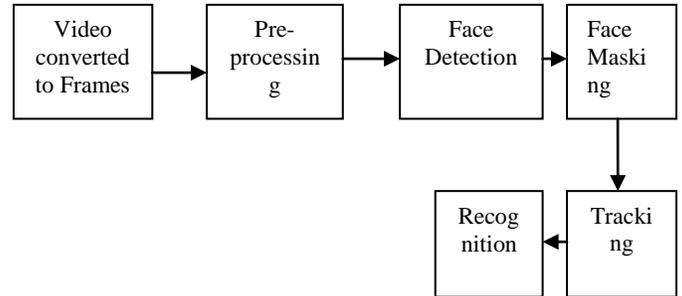


Fig.1. System Architecture of Detecting and Recognizing Face Images from Videos using Spherical Harmonics and RBF kernel Techniques

B. Recognizing Images across Obscure and Various Lighting Conditions using LBP Feature Extraction and SVM Classification Techniques

To recognize images across Obscure and Various lighting Conditions we sharpen and enhance the blur image with different lighting conditions which is a test image and images in database for performing pre-processing using Gaussian filter. And then we extract features using LBP feature extractor. We calculate the weighted mapping for the test and the dataset images. After the feature set has been computed for each pixel, this feature will be passed to the SVM classifier, the classification stage has two components, a training phase and a testing phase. In training phase features of images in dataset are classified in to different classes and in testing phase based on the features of test images classifier decides to which class of the dataset the input image belongs to.

Modules for Recognizing Images across Obscure and Various Lighting Conditions using LBP Feature Extraction and SVM Classification Techniques

- Pre-processing
- Normalization
- Feature Extraction
- Recognition

Pre-processing

In pre-processing method we use Gaussian filter to remove unwanted noise from the input images. The Gaussian filter is a nonlinear digital filtering technique normally used in removing noise. Noise reduction is an important step in any recognition techniques. Gaussian filter perform smoothing of the image without having more affects on edges of image which are the major characteristic of an image i.e. it preserves the edges even after smoothing.

Normalization

Normalization is a process that changes the range of pixel intensity values. Illumination variation caused by changes in sources of light at different positions and various intensities causes large variations. To overcome this problem we obtained a new method of performing image normalization. This method removes shadows and specularities from images. All shadowed regions are given a uniform colour and then it eliminates the soft shadows and specularities and thus creates an illumination invariant copy of the original image.

LBP Feature Extraction

LBP is a texture descriptor which can be used to recognize faces. This operator assigns label to each pixel of image. Taking centre pixel as reference it assigns binary values to pixels surrounding it. It assigns value 1 if the pixel value is greater than centre pixel value else value 0. Combining these 0s and 1s we get a binary value. We convert this binary value to a decimal number and the resulting decimal number is given as label to the centre pixel. This process is repeated throughout the image. Later based on these labels classification is performed.

Recognition

SVM classifier is used to recognize images. SVM classifier classifies two classes by drawing a hyper plane between two classes. The hyper plane is drawn such that it is at a equal distance between the features of both class which is close to other group and must have a maximum distance between the features. Let us assume that first the classifier classifies two classes as class 1 and class 2 by drawing a hyper plane between them. Now if a new test case is entered this classifier decides which class does this test sample lies by observing the feature placement. If majority of features of new test sample lies on one side i.e. class 1 side of hyper plane then this test sample belongs to class 1. If it lies on other side i.e. class 2 of hyper plane then it belongs to class 2. After classifying we analyze about our classification. We analyze if there is any misclassification in our result. Then we find accuracy of our system. The system architecture for recognizing images across obscure and various lighting conditions using LBP feature extraction and SVM classification techniques is shown in Fig.2.

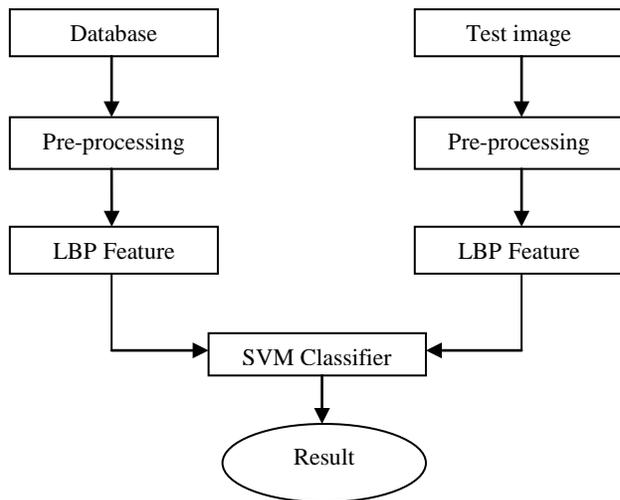


Fig.2. System Architecture for Recognizing Images across Obscure and Various Lighting Conditions using LBP Feature Extraction and SVM Classification Techniques

IV. RESULT AND DISCUSSIONS

A. Detecting and Recognizing Face Images from Videos using Spherical Harmonics and RBF kernel Techniques

Detecting and Recognizing face images from videos is tested using three different videos. This system has four folders, they are – Dataset, Face images, Frames and Pre-processing.

Dataset is where three videos are saved.

Face images where the cropped images of some (10) frames of each video is placed.

Frames where frames obtained after frame conversion get stored.

Pre-processing where video after Pre-processing gets stored.

- Run the main code.
- There pops up a GUI having 3 radio buttons one for each video and 2 other buttons for Frame conversion and preprocessing.
- Choose any one video
- Once the video starts playing choose frame conversion. Here video will be converted to frames and all frames are stored in folder Frames.
- Once frame conversion is over select pre-processing. Now the video will be preprocessed and the pre-processed video will be stored in Pre-processor folder.
- Once the pre-processing is completed. Second GUI pops up automatically.
- This second GUI will have four buttons. They are – Face Detection, Gradient Map, Histogram Calculation and Tracking.
- When Face Detection is pressed the face image from the video is detected having a yellow box around it.
- When gradient Map is pressed the gradient map of video starts playing till the end of video.
- Later Histogram Calculation button is pressed. Now we get the Histograms of RGB separately.
- Finally when Tracking is pressed the video starts playing with a red circle around the face image of a person in video till the end.

- After the completion of tracking the third GUI will pop up automatically. It has for buttons they are – Face Recognition, Performance Measurement, Clear and Close.
- When Face Recognition is pressed the Recognized image is displayed.
- When Performance Measurement is pressed the performance graphs are obtained.
- When clear and close is pressed everything gets cleared i.e. All the GUIs and all the pop ups and figures get closed.

Step by step implementation for Detecting and Recognizing Face Images from Videos using Spherical Harmonics and RBF kernel Techniques is shown in Fig.3.

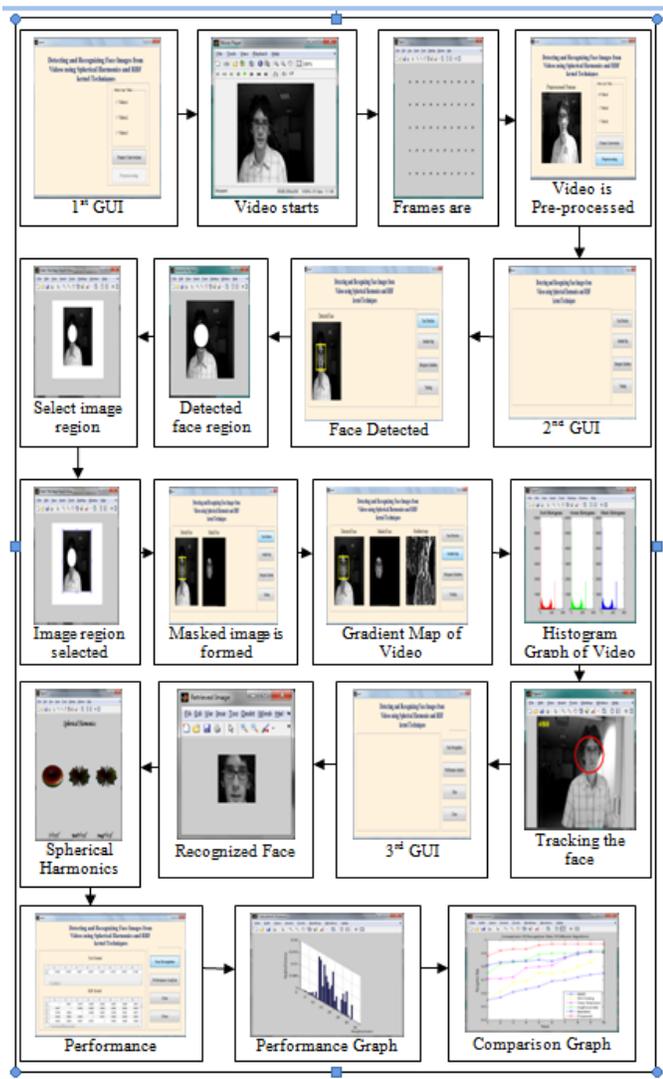


Fig.3. Detecting and Recognizing Face Images from Videos using Spherical Harmonics and RBF kernel Techniques

A comparative study of various face detection and recognition techniques across videos on three different videos is tabulated in Table 1.

Table.1. Face Recognition Rate across videos

Author	Method	Database	Recognition Rate (%)
Pojala Chiranjeevi, <i>et al</i> [21]	Personalized Appearance Models	CK+	75%
Chao Xiong, <i>et al</i> [22]	Gabor Feature, LBP Feature, SIFT (Scale Invariant feature Transform) and SVM	CK+	68.87%
Sujitha Martin, <i>et al</i> [23]	Automatic Machine vision, Gaze zone estimation	CK+	71%
Himanshu S. Bhatt, <i>et al</i> [24]	Clustering, Re-Ranking and Fusion	CK+	23.8%
Spherical Harmonics and RBF kernel Techniques (Proposed system)		CK+	98%

Table 1 indicates the proposed Detecting and Recognizing Face Images from Videos using Spherical Harmonics and RBF kernel Techniques gives the best recognition rate.

B. Recognizing Images across Obscure and Various Lighting Conditions using LBP Feature Extraction and SVM classification Techniques

Recognizing faces across obscure and varying light conditions is tested using Yale b database which consists of illuminated images. This system has two folders image and train image. Image where test (input) images are stored and train image where illuminated images from Yale B database are stored

The steps to be followed when we run the Train_main.m code are:

- A GUI pops up with 2 buttons Train Database and Next. Press Train database.
- All train images will be loaded now press next.
- A GUI pops up having 6 buttons. They are – Input Image, Pre-processing, Normalization, LBP feature, Recognition, Analysis.
- When Input image is pressed a pop up window of folders comes where we need to select the input image. This image will be displayed on GUI.
- When Preprocessing is pressed this test image is preprocessed and the preprocessed image is displayed on the GUI.
- When Normalization is pressed the preprocessed image is normalized, and the normalized image is displayed on GUI.
- When LBP Feature is pressed the LBP features of the image is calculated and the resulting table is displayed on GUI.
- When Recognition is pressed the recognized image from the training set is displayed.
- Finally When Analysis is pressed the performance graph is plotted.

Step by step implementation for Recognizing Faces across Obscure and Various Lighting Conditions using LBP Feature Extraction and SVM classification Techniques.

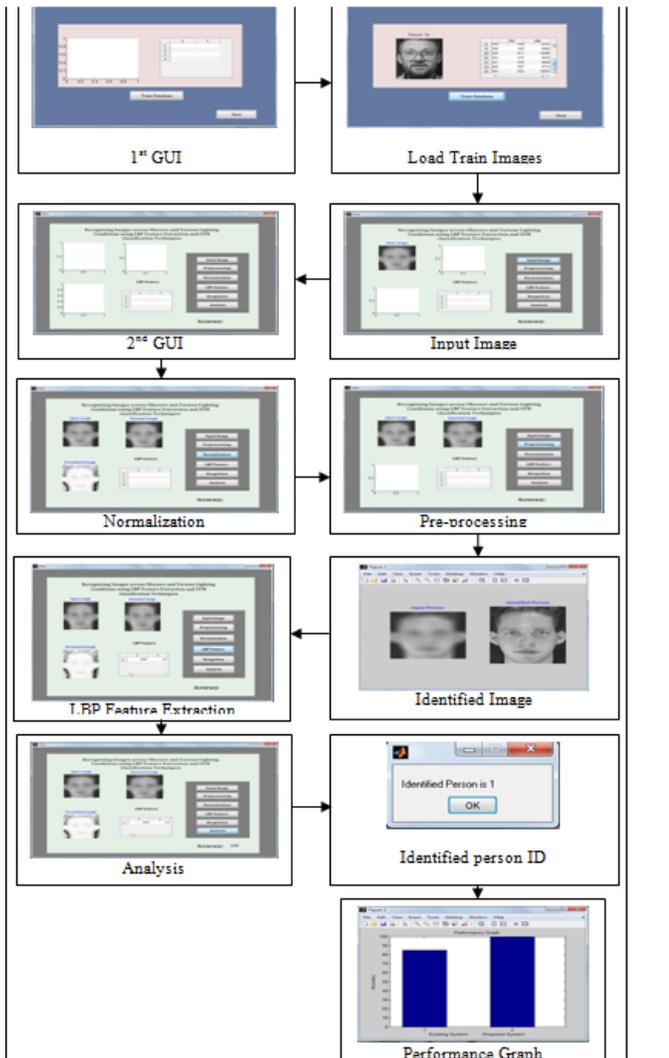


Fig.4. Recognizing Images across Obscure and Various Lighting Conditions using LBP Feature Extraction and SVM classification Techniques

A comparative study of various face recognition techniques under obscure and various lighting conditions on Yale B database is tabulated in Table 2.

Table.2. Face Recognition Rate under Obscure and Various Lighting Conditions

Author	Method	Database	Recognition Rate (%)
Abhijith Punnappurath, <i>et al</i> [8]	TSF (Trading Support Facility) model	Yale B	76.27%
Chi Ho Chan, <i>et al</i> [25]	Multiphase Local Phase Quantization (LPQ), kernel	Yale B	87%

	fusion of multiple descriptors		
Zhengwu Zhang, <i>et al</i> [26]	Riemannian Framework	Yale B	65%
Zeynep Yucel, <i>et al</i> [27]	Gaussian Process Regression, Neural Networks, Saliency Schemes	Videos	80%
LBP Feature Extraction and SVM Classification (Proposed System)		Yale B	100%

Table 2 indicates the proposed Face Recognition Technique using LBP Feature Extraction and SVM Classification gives best recognition rate on Yale B database.

V. CONCLUSION

In this paper we have developed two novel techniques to detect and recognize faces from Videos and across Obscure and Various Lighting Conditions. Detecting and recognizing face images from Videos is proposed using Spherical Harmonics and Radial Basis Function (RBF) kernel Technique and tested using three different videos. Recognizing images across Obscure and Various Lighting Conditions is proposed using Local Binary Pattern (LBP) Feature Extraction and Support Vector Machine (SVM) Classification Techniques and tested using Yale B database which consists of images across various lighting conditions. Both the proposed systems were tested across various existing state of the art techniques. From our analysis we have found that proposed detecting and recognizing faces from videos using Spherical Harmonics and RBF kernel Techniques gives the best recognition rate of 98% and Recognizing faces across Obscure and Various Lighting Conditions using LBP Feature Extraction and SVM Classification gives the best recognition rate of 100% on Yale B database

REFERENCES

- [1] V. Blanz., T. Grother., P. J. Phillips., T. Vetter.: Face recognition based on frontal views generated from non-frontal images, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Vol. 2 (2005)
- [2] P. Breuer., K.-I. Kim., W. Kienzle., B. Scholkopf., V. Blanz.: Automatic 3D face reconstruction from single images or video, in Proc. IEEE Int. Conf. Autom. Face Gesture Recognit, No. 160 (2008)
- [3] Shunli Zhang, Xin Yu., Yao Sui, Sicong Zhao., Li Zhang.: Object Tracking with Multi-View Support Vector Machines. IEEE Transactions on Multimedia, Vol. 17, No. 3 (2015)
- [4] Derpanis. K. G., Sizintsev. M, Cannons. K. J., Wildes. R. P.: Action Spotting and Recognition Based on a Spatiotemporal Orientation Analysis. IEEE Transactionson Pattern Analysis and Machine Intelligence, Vol. 35, No. 3 (2013)
- [5] Shunli Zhang, Xin Yu., Yao Sui, Sicong Zhao., Li Zhang.: Object Tracking with Multi-View Support Vector Machines. IEEE Transactions on Multimedia, Vol. 17, No. 3 (2015)

- [6] Derpanis. K. G., Sizintsev. M, Cannons. K. J., Wildes. R. P.: Action Spotting and Recognition Based on a Spatiotemporal Orientation Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 3 (2013)
- [7] Steven L. Fernandes., G. Josemin Bala.: 3D and 4D Face Recognition: A Comprehensive Review. *Recent Patents on Engineering*, Vol. 8, No. 2 (2014)
- [8] Abhijith Punnapurah., Ambasamudram Narayanan Rajagopalan., Sima Taheri., Rama Chellappa., Guna Seetharaman.: Face Recognition Across Non-Uniform Motion Blur, Illumination, and Pose. *IEEE Transactions on Image Processing*, Vol. 24, No. 7 (2015)
- [9] Taegeun Oh., Sanghoon Lee.: Blind Sharpness Prediction Based on Image-Based Motion Blur Analysis. *IEEE Transactions on Broadcasting*, Vol. 61, No. 1 (2015)
- [10] Priyanka Vageeswaran., Kaushik Mitra., Rama Chellappa.: Blur and Illumination Robust Face Recognition via Set-Theoretic Characterization. *IEEE Transactions on Image Processing*, Vol. 22, No. 4 (2013)
- [11] Steven L. Fernandes., G. Josemin Bala., P. Nagabhushan., S. K Mandal.: A Comparative Study on Score Level Fusion Techniques and MACE Gabor Filters for Face Recognition in the presence of Noises and Blurring effects. *IEEE Int. Conf. on Cloud & Ubiquitous Computing and Emerging Technologies* (2013)
- [12] Steven L. Fernandes., G. Josemin Bala.: A Comparative Study On ICA and LPP Based Face Recognition Under Varying Illuminations And Facial Expressions. *IEEE Int. Conf. on Signal Processing, Image Processing and Pattern Recognition* (2013)
- [13] Mohammad Reza Faraji., Xiaojun.: Face Recognition under Varying Illumination with Logarithmic Fractal Analysis. *IEEE Signal Processing Letters*, Vol. 21, No. 12 (2014)
- [14] Bing Li., Weihua Xiong., Weiming Hu., Brian Funt.: Evaluating Combinational Illumination Estimation Methods on Real-World Images. *IEEE Transactions on Image Processing*, Vol. 23, No. 3 (2014)
- [15] Shida Beigpour., Christian Riess., Joost van de Weijer., Elli Angelopoulou.: Multi-Illuminant Estimation With Conditional Random Fields. *IEEE Transactions on Image Processing*, Vol. 23, No. 1 (2014)
- [16] Cong Phuoc Huynh., Antonio Robles-Kelly.: Segmentation and Estimation of Spatially Varying Illumination. *IEEE Transactions on Image Processing*, Vol. 23, No. 8 (2014)
- [17] An Jin., Birsen Yazıcı., Vasilis Ntziachristos.: Light Illumination and Detection Patterns for Fluorescence Diffuse Optical Tomography Based on Compressive Sensing. *IEEE Transactions on Image Processing*, Vol. 23, No. 6 (2014)
- [18] Yanli Wan., Zhenjiang Miao., Xiao-Ping Zhang., Zhen Tang., Zhifei Wang.: Illumination Robust Video Foreground Prediction Based on Color Recovering. *IEEE Transactions on Multimedia*, Vol. 16, No. 3 (2014)
- [19] Steven L. Fernandes., G. Josemin Bala.: Recognizing Faces When Images Are Corrupted By Varying Degree of Noises and Blurring Effects. *Advances in Intelligent Systems and Computing*, Vol. 337, No. 1 (2015)
- [20] Chia-Ping Chen., Chu-Song Chen.: Intrinsic Illumination Subspace for Lighting Insensitive Face Recognition. *IEEE Trans. on Systems, Man and Cybernetics- part B: Cybernetics*, Vol. 42, No. 2 (2012)
- [21] Pojala Chiranjeevi., Viswanath Gopalakrishnan., Pratibha Moogi.: Neutral face Classification Using Personalized Appearance Models for Fast and Robust Emotion Detection. *IEEE Transactions on Image Processing*, Vol. 24, No. 9 (2015)
- [22] Chao Xiong., Guangyu Gao., Zhengjun Zha., Shuicheng Yan., Huadong Ma., Tae-Kyun Kim.: Adaptive Learning for Celebrity Identification with video Context. *IEEE Transactions on Multimedia*, Vol. 16, No. 5 (2014)
- [23] Sujitha Martin., Ashish Tawari., Mohan Manubhai.: Toward privacy-Protecting Safety Systems for Naturalistic Driving Videos. *IEEE Transactions on Intelligent transportation Systems*, Vol. 15, No. 4 (2014)
- [24] Himanshu S. Bhatt, Richa Singh., Mayank Vatsa.: On Recognizing Faces in Videos using Clustering-Based Re-Ranking and Fusion. *IEEE Transactions on Information Forensics and Security*. Vol. 9, No. 7 (2014)
- [25] Chi Ho Chan., Muhammad Atif Tahir., Josef Kittler., Matti Pietikainen.: Multiscale Local Phase Quantization for Robust Component-Based Face Recognition Using Kernel Fusion of Multiple Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 5 (2013)
- [26] Zhengwu Zhang., Eric Klassen., Anuj Srivastava.: Gaussian Blurring-Invariant Comparison of Signals and Images. *IEEE Transactions on Image Processing*, Vol. 22, No. 8 (2013)
- [27] Zeynep Yucel., Albert Ali Salah., Cetin Mericli., Tekin Mericli., Roberto Valenti., Theo Gevers.: Joint Attention by Gaze Interpolation and Saliency. *IEEE Transactions on Cybernetics*, Vol. 43, No. 3 (2013)

A Comparative Study to Recognize Surgically Altered Images

Steven Lawrence Fernandes, Dr. G. Josemin Bala

Abstract—In this paper we have developed two novel methods to recognize faces across surgically altered faces using two techniques. Recognizing faces across surgically altered using LBP and LDP techniques and tested using IIIT-Delhi Plastic Surgery Face Database which consists of images across varying surgically altered faces. Recognizing faces under LBP and PCA techniques and tested using IIIT-Delhi Plastic Surgery Face Database. Both the proposed systems were tested across various existing state of the art techniques. From our analysis we have found that proposed recognizing faces across surgically altered using LBP and LDP technique gives the best recognition rate of 81.81% on IIIT-Delhi Plastic Surgery Face Database and using LBP and PCA technique gives the best recognition rate of 87.5% on combined heterogeneous face database.

Keywords—Face recognition; Local Binary Pattern; Local Derivative Pattern; and Principle Component Analysis.

I. INTRODUCTION

Identifying human faces across image processing is a difficult task mainly when it comes to surgically altered faces. We all know that surgically altered faces can cause drastic changes in the human face. It is difficult for the human visual system to identify such altered faces. So in this paper we are trying to overcome this difficulty by proposing two methodologies. They are - Recognizing surgically altered faces using LBP (Local Binary Pattern) and LDP (Local Derivative Pattern) techniques and Recognizing surgically altered faces using LBP and PCA (Principle Component Analysis) Techniques [1][2].

II. RELATED WORK

Here we have done some study of previous works on surgically altered faces which is helpful in knowing more about the problems and various techniques to solve them. Some of the study contents are listed below.

Author in [3] have proposed that in many areas of research and industrial situations, including many data analytic problems in chemistry, a strong nonlinear relation between different sets of data may exist. While linear models may be a good simple approximation to these problems, when nonlinearity is severe they often perform unacceptably. The nonlinear Partial Least Squares (PLS) method was developed in the area of chemical data analysis. A specific feature of PLS is that relations between sets of observed variables are

modeled by means of latent variables usually not directly observed and measured. Since its introduction, two methodologically different concepts of fitting existing nonlinear relationships initiated development of a series of different nonlinear PLS models.

Authors in [4] have proposed a method called dual-attributes for face verification, which is robust to facial appearance changes caused by cosmetics or makeup. Attribute-based methods have shown successful applications in a couple of computer vision problems, e.g., object recognition and face verification. However, no previous approach has specifically addressed the problem of facial cosmetics using attributes. Our key idea is that the dual-attributes can be learned from surgically altered faces, separately. Then the shared attribute scan is used to measure facial similarity irrespective of surgical changes. In essence, dual-attributes are capable of matching faces before and after surgery in a semantic level, rather than a direct matching with low-level features.

III. PROPOSED SYSTEM

In this paper we have worked on two difficulties in recognizing faces and have proposed two systems to work on these problems having one methodology for each problem. The proposed two methodologies are explained below.

A. Recognizing surgically altered faces using LBP and LDP

Here we proposed a novel method to recognize the surgical images. We extract the feature of pre-surgical and post surgical images. For feature Extraction we will use the LBP and LDP. After Extracting feature, it will be matching with the pre-surgical image for classification. From that LDP feature we extract the histogram values. This histogram values will be used to mapping the dataset feature. Finally it recognize from the database.

Modules for Recognizing surgically altered faces using LBP and LDP

- Face granulation
- LBP feature Extraction and LDP feature extraction
- Histogram matching and Micro Pattern Generation

Face granulation:

In face granulation first we segment the face from the image. For segmenting face, we convert the image as ycbcr. We

detect the skin region from the image. Then we apply the morphological operation to extract the face from the image.

LBP and LDP feature Extraction:

Initially we separate the image as patches. For each patch of image we apply the LBP.

The n th-order LDP can be encoded by $(n-1)$ th order local derivative various, to calculate second-order LDP must calculate first-order derivative.

Histogram and Micro pattern Generation:

First we calculate the first derivation of each pixel. And then, we can calculate the multiple of the first derivations between operating pixel and its neighbors. In 0° , we will get $LDP\ 0^\circ = "01010100"$. As the same, we will get $LDP45^\circ = "00101111"$, $LDP90^\circ = "11010000"$ and $LDP135^\circ = "11000110"$. The corresponding histograms were displayed.

System Architecture for recognizing surgically altered faces using LBP and LDP

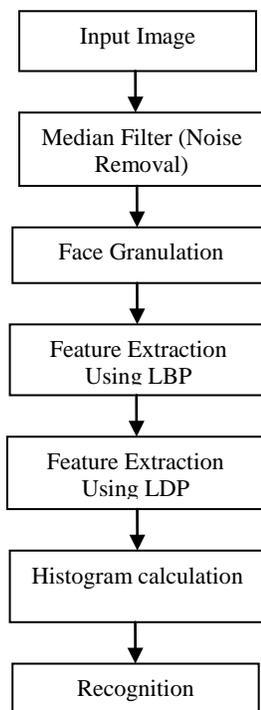


Fig: 1. System Architecture for recognizing surgically altered faces using LBP and LDP

B. Recognizing surgically altered faces using LBP and LDP

Here we proposed a novel method to recognize the surgical images. We extract the feature of pre-surgical and post surgical images. For feature Extraction we will use the LBP and PCA. After Extracting feature, it will be matching with the pre-surgical image for classification. From that PCA feature we extract the histogram values. This histogram values

will be used to mapping the dataset feature. Finally it recognize from the database.

Modules for Surgically altered faces using LBP and PCA Techniques:

- Preprocessing.
- Feature Extraction.
- Classification
- Performance Analysis.

Preprocessing:

The images were preprocessed by resizing and denoising. The input images were resized to a common size 256x256. The denoising is done using median filter which is a common filter employed for all the preprocessing steps. The median filter identifies the noisy pixels in the images and replaces the noisy pixels based on the median value of the neighboring pixels.

Feature Extraction:

The face images were separated into patches. The features were extracted from the separated patches of the input face images. The LBP and PCA features were extracted for each patch. The statistical features such as mean, standard deviation and entropy were calculated. The calculated features were arranged and saved.

Classification:

The input images were identified whether the face of a person in the train surgically altered or not using (Support Vector Machine) SVM classifier. The SVM classifier trains and arranges the feature vectors in hyper planes. The virtual hyper plane separates the two categories of the images i.e. Test and Train. Now if a test image is given based on the number of features relied on it decides which class this image belongs to. If majority of features lie on the side of hyper plane where images with surgically altered relies then the image belongs to test class. Otherwise if majority of feature lies on other side of the hyper plane then it belongs to the class train.

Performance Analysis:

The performance of the system is measured by calculating the accuracy, Sensitivity and specificity of the classifier. The accuracy of the classifier represents to which extend the classifier classifies the images based on the given label. The sensitivity of the classifier represents how exactly the classifier correctly classifies the data to each category. The specificity of the classifier represents how exactly the classifier correctly rejects the data to each category.

Recognizing surgically altered faces using LBP and LDP

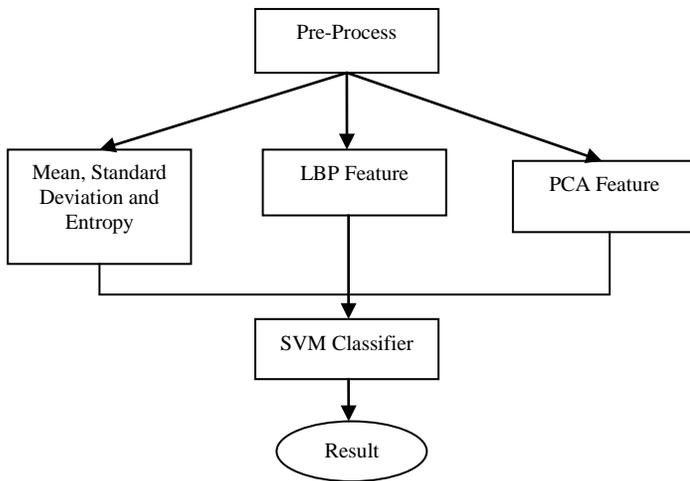


Fig. 2. System Architecture for surgically altered faces using LBP and PCA Technique

IV. RESULT AND DISCUSSIONS

A. Recognizing surgically altered faces using LBP and LDP

Recognizing surgically altered faces are tested using IIIT-Delhi Plastic Surgery Face Database which consists of images across varying surgically altered faces. In our system train is done using 1 Pre-surgery and test is done using 1 Post-surgery image.

The steps to be followed when we run the main code (main.m) are:

- A Main window will open which is the guide design of our project.
- First we have to provide an input image, we have 4 surgically altered data and selecting 1 image.
- Next we have to press pre-processing, it will display the denoised image.
- When we press the face segmentation, face part only segmented from the original image. That will be displayed in the second axis.
- In face granulation we extracting the important parts of the face such as eyes, nose and lips of the segmented image.
- Next extracting the LBP feature of each portion such as eyes, nose and lips and corresponding LBP Histogram will appear.
- Next compute the LDP feature it will display in the third axis. This feature we are extracting the histogram values each histogram value is map to the data base images. Finally the recognized image will be display.

Step by step implementation for Recognizing surgically altered faces using LBP and LDP is shown in Fig.3

A comparative study of various face recognition techniques across surgically altered faces on IIIT-Delhi Plastic Surgery Face Database is tabulated in Table 1.

Table: 1. Face Recognition Rate across surgically altered faces

Author	Method	Database	Recognition Rate (%)
Minal Mun et al.[18]	PCA,LBP	Look Alike	78.27
Bincy Baby et al.[19]	PCA,EUCLBP,SIFT	Look Alike	83.33
Vaibhav R Wagare et al. [20]	PCA, LBP	Look Alike	69.58
	LBP,LDP,PCA	IIIT-Delhi Plastic Surgery	87.5

Table 1 indicates the proposed Recognizing Faces across surgically altered faces using LBP and LDP Technique gives the best recognition rate on IIIT-Delhi Plastic Surgery Face Database.

B. Recognizing surgically altered faces using LBP and PCA

Recognizing surgically altered faces are tested using IIIT-Delhi Plastic Surgery Face Database which consists of images across varying surgically altered faces. In our system train is done using 1 Pre-surgery and test is done using 1 Post-surgery image.

The steps to be followed when we run the main code (main.m) are:

- A Main window pops up with 6 buttons. They are – Input Image, Preprocessing, Feature Extraction, Makeup Detection, Face Recognition, and Performance.
- When the Input image is pressed a window pops up asking to select the image before surgery, when we chosen the image it will load to the window. Later we have to load the image of after surgery it will also load to the window.
- When Preprocessing is pressed the preprocessed image gets displayed on the window.
- When Feature Extraction is pressed, Face Patches of both images will be displayed in different window along with Feature values of both images will be displayed in separate window.
- When Makeup Detection is pressed, It will detect in both the images, if makeup detected window will open by saying that makeup is detected in the first else no makeup detected in the first image. It will detect for both the images.
- When we press the Face Recognition, if the persons were same it show a window by saying that two persons were same else two persons were different.
- When Performance is pressed, correlation feature points, confusion matrix, ROC and Recognition rate of both the image will be displayed.

Step by step implementation for Recognizing surgically altered faces using LBP and PCA is shown in Fig.4



Fig. 3. recognizing surgically altered faces using LBP and LDP Technique



Fig: 4. Recognizing surgically altered faces using LBP and PCA Techniques

V. CONCLUSION

In this paper we have developed two novel methods to recognize faces across surgically altered faces using two techniques. Recognizing faces across surgically altered using LBP and LDP techniques and tested using IIIT-Delhi Plastic Surgery Face Database which consists of images across varying surgically altered faces. Recognizing faces under LBP and PCA techniques and tested using plastic surgery face databases combined heterogeneous face database. Both the proposed systems were tested across various existing state of the art techniques. From our analysis we have found that proposed recognizing faces across surgically altered using LBP and LDP Technique gives the best recognition rate of 81.81% on IIIT-Delhi Plastic Surgery Face Database and using LBP and PCA technique gives the best recognition rate of 87.5% on combined heterogeneous face database.

dimension symmetrical image matrix, 24th Chinese Control and Decision Conference (CCDC), (May 2012)

REFERENCES

- [1] Yu Su., Shiguang., Xilin Chen., WEn Gao.: Hierarchical Ensemble of Global and Local Classifiers for Face Recognition. *Computer Vision*, 2007. ICCV 2007. IEEE 11th International conference. pp 1--8, 14--21 (Oct. 2007)
- [2] Liu Z., Yang J., Liu C.: Extracting multiple features in the CID color space for face recognition. *IEEE Trans Image Process* 19(9), pp.2502--2509 (2010)
- [3] Roman Rosipal.: Nonlinear Partial Least Squares An Overview. Medical University of Vienna, Austria and Pacific Development and Technology. pp.21 (2011)
- [4] Guodong Guo., Lingyun Wen., Shuicheng Ya.: Face Authentication with Make up Changes. *Circuits and Systems for Video Technology*, IEEE Transaction. Vol.24, pp.814--825 (2014)
- [5] Xiaoli Li and Feipeng Da.: "Robust 3D Face Recognition Based on Rejection and Adaptive Region Selection", ACCV'09 Proceedings of the 9th asian conference on Computer Vision, Vol. Part III, pp 581--590.
- [6] Hashmat Popat.: Determining normal and abnormal lip shapes during movement for use as a surgical outcome measure, School of dentistry, Cardiff university, (Aug. 2012)
- [7] Verma T., CSIT Durg., Durg., Sahu R.K.: PCA-LDA based face recognition system & results comparison by various classification techniques, IEEE International Conference on Green High Performance Computing (ICGHPC), (Mar. 2013)
- [8] Hsi-Kuan Chen., Yi-Chun Lee., Chin-Hsing Chen.: Gabor feature based classification using Enhance Two-direction Variation of 2DPCA discriminant analysis for face verification, IEEE International Symposium on Next-Generation Electronics (ISNE), (Feb. 2013)
- [9] Fernandes S.L., Bala G.J.: A comparative study on ICA and LPP based Face Recognition under varying illuminations and facial expressions, International Conference on Signal Processing Image Processing & Pattern Recognition (ICSIPR), (Feb. 2013)
- [10] Shen Maodong., Cao Jiangtao., Li Ping.: Independent component analysis for face recognition based on two

Extraction of Blood Vessels and Optic Disc Segmentation for Retinal Disease Classification

Jestin.V.K

Assistant professor

Dept of ECE,

Hindusthan College of Engineering and Technology,

Coimbatore, Tamilnadu.

vjestin8@gmail.com

Abstract— The retina is the only tissue in human body from which the information of blood vessel can be unswervingly obtained. The information of retinal vessel plays an important role in the diagnosis and treatment of various diseases such as glaucoma, age-related macular degeneration, degenerative myopia, diabetic retinopathy etc. The morphology of the retinal blood vessel and the optic disc is an important structural indicator for assessing the presence and severity of retinal diseases such as diabetic retinopathy, hypertension, glaucoma, hemorrhages, vein occlusion, and neovascularization. However, to assess the diameter and tortuosity of the retinal blood vessel or the shape of the optic disc, manual planimetry has commonly been used by ophthalmologists, which is generally time consuming and prone to human error, especially when the vessel structures are complicated or a large number of images are acquired to be labeled by hand. Therefore, a reliable automated method for retinal blood vessel and optic disc segmentation, which preserves various vessel and optic disc characteristics, is attractive in computer-aided diagnosis. However here implement a new competent method for the detection diseases using the retinal fundus image. In this anticipated work first step is the extraction of retinal vascular tree using graph cut technique. The blood vessel information is then use to calculate approximately the position of optic disc. These results are given to an ANN classifier for the detection and classification of diseases. By robotically identify the disease from normal images, the workload and its costs will be reduced.

Keywords—Fundus Images, optic disc, Diabetic Retinopathy, Hypertension, Glaucoma, ANN.

I. INTRODUCTION

CSCC will do the final formatting of your paper. If your paper is intended for a conference, please observe the conference page limits. Retinal image analysis is one of the crucial topics in medical image processing. During the last few centuries, people are trying to extract the various features like blood vessels, optic disc, macula, fovea are automatically from retinal image. [1]. The Fundus Image Analysis system described in this paper is developed to assist ophthalmologist's diagnosis by providing second opinion and also functions as an automatic tool for the mass screening of diabetic retinopathy. Color fundus images are used by ophthalmologists to study eye diseases like Diabetic Retinopathy(DR), Age related Macular Degeneration (AMD) hypertension, glaucoma, vein occlusion and Retinopathy of pre-maturity (ROP). Extraction of the normal features like optic disc, fovea, blood vessels and abnormal features like exudates, cotton wool spots,

Microaneurysms (MA) and hemorrhages from color fundus images are used in fundus image analysis system for comprehensive analysis and grading of Diabetic Retinopathy (DR) [2].

For the diagnosis of complete diseases, assessment of retinal blood vessel is significant. It offer a lot information conversely for easy recognition of exudates or microaneurysms [3]. Diabetic retinopathy is cause by mutually the forms of diabetes i.e. diabetes mellitus and diabetes insipidus. It is a extremely asymptomatic disease in the premature stages and it could lead to lasting vision loss if untreated for long time. The problem here is the patients may not know about it until it reaches advanced stages. Once it reach advanced stages vision loss become unavoidable [4].

Here we using graph cut technique for blood vessel segmentation. we have implemented a preprocessing method, which consists of an efficient adaptive histogram equalization and robust distance transform. This operation improves the robustness and the accuracy of the graph cut algorithm.[5] The optic disc segmentation starts by defining the location of the optic disc. This method used the convergence features of vessels into the optic disc to calculate approximately its location. The disc area is then segmented using two different automated methods (MRF image reconstruction and compensation factor). Both methods use the convergence feature of the vessels to identify the position of the disc. [6] .

The purpose of image classification scheme is to assign each input to one of the diseases pattern classes. It is the process of assigning a label to each unknown input image. In this work, the artificial neural network approach namely, Back propagation network (BPNs) is used to classify the images The back propagation algorithm is used in layered feed-forward ANNs. This means that the synthetic neurons are prearranged in layer and drive their signal "forward", and then the errors are propagated backwards. The system receives input by neurons in the input layer, and the output of the system is given by the neurons on an output layer.[7] Glaucoma is a term describing a group of ocular (eye) disorders that result in optic nerve damage, often associated with increased fluid pressure in the eye (intraocular pressure)(IOP). The disorders can be roughly divided into two main categories, "open-angle" and "closed-angle glaucoma.[9]

Vision is the most advanced human sense. So images play the most important role in human perception. The human eye is nearly in the shape of a sphere. Its average diameter is approximately 20 mm. The eye is made up of three coats,

enclose three apparent structures. The outermost layer is composed of the cornea and sclera. [10]

The morphology of the retinal blood vessel and the optic disk is an important structural indicator for assessing the presence and severity of retinal diseases such as diabetic retinopathy, hypertension, glaucoma, hemorrhages, vein occlusion, and neovascularization. However, to assess the diameter and tortuosity of the retinal blood vessel or the shape of the optic disk, manual planimetry has commonly been used by ophthalmologists, which is generally time consuming and prone to human error, especially when the vessel structures are complicated or a large number of images are acquired to be labeled by hand. Therefore, a reliable computerized method for retinal blood vessel and optic disk segmentation, which preserves various vessel and optic disk characteristics, is attractive in automatic diagnosis. [11]

Blood vessels can be seen as thin elongated structures in the retina, with variation in width and length. In order to segment the blood vessel from the fundus retinal image, we have implemented a preprocessing technique, which consists of an effective adaptive histogram equalization and robust distance transform. This operation improves the robustness and the accuracy of the graph cut algorithm. [12]

II. METHODOLOGY

The proposed method is made up of four fundamental parts. Basic system level block diagram is shown below:

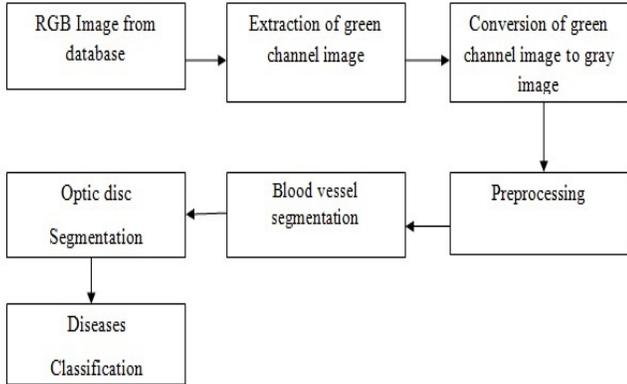


Fig 1: Basic blocks

A. Blood Vessel Segmentation

Blood vessels can be seen as thin lengthened structure in the retina, with variation in width and length. In order to fragment the blood vessel from the fundus retinal image, we have implemented a preprocessing method, which consists of an efficient adaptive histogram equalization and robust distance transform. This action improves the robustness and the accuracy of the graph cut algorithm. Fig. 2 shows the illustration of the vessel segmentation algorithm.[6]

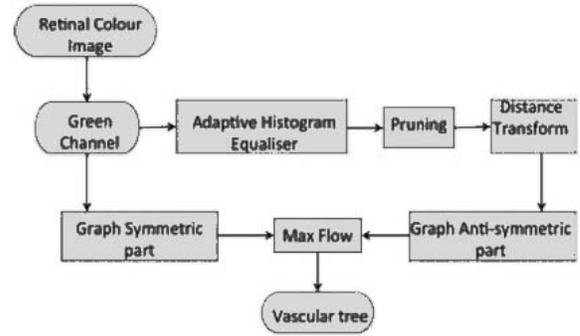


Fig. 2. Vessel segmentation algorithm.

The graph cut is an energy-based object segmentation approach. The method is characterized by an optimization process designed to reduce the power generated from a given image data. This power defines the relationship between neighborhood pixel elements in an image. The graph cut method is used in our detection because it allows the amalgamation of prior knowledge into the graph formulation in order to guide the model and find the optimal segmentation.[6]

To concentrate on the beyond mentioned trouble, the segmentation of blood vessels using the graph cut need a special graph formulation. One of the method used to deal with the shrinking bias problem is to require an extra connectivity prior, where the user marks the restraint connectivity [8]. In order to attain full automatic segmentation, we used the method presented in [9], which overcomes the “shrinking bias” by adding the mechanism of vectors flux into the production of the graph.

B. Graph Construction for Vessel Segmentation

The graph cut is an energy-based object segmentation approach. The technique is characterized by an optimization operation designed to minimize the energy generated from a given image data. This energy defines the relationship between neighborhood pixel elements in an image. A graph $G(v, \epsilon)$ is defined as a set of nodes (pixels) v and a set of undirected edges ϵ that connect these neighboring nodes. The graph included two special nodes, a foreground (Fg) terminal (source S) and a Bg terminal (sink T). ϵ includes two types of undirected edges: neighborhood links (n-links) and terminal links (t-links). Each pixel $p \in P$ (a set of pixels) in the graph presents two t-links $\{p, S\}$ and $\{p, T\}$ connecting it to each terminal, while a pair of neighboring pixels $\{p, q\} \in N$ (number of pixel neighbors) is connected by an n-link [13]. Thus,

$$\epsilon = N \bigcup_{p \in P} \{\{p, S\}, \{p, T\}, \nu = P \cup \{S, T\}\} \dots \dots \dots (1)$$

The graph cut technique is used in our segmentation because it allows the incorporation of prior knowledge into the graph formulation in order to guide the model and find the optimal segmentation. Let us assume $A = (A1, Ap, \dots, AP)$ is a binary vector set of labels assigned to each pixel p in the image,

where A_p indicate assignments to pixels p in P . Therefore, each assignment A_p is either in the F_g or B_g . Thus, the segmentation is obtained by the binary vector A and the constraints imposed on the regional and boundary proprieties of vector A are derived by the energy formulation of the graph defined as

$$E(A) = \lambda \cdot R(A) + B(A) \dots\dots\dots(2)$$

Where the positive coefficient λ indicates the relative importance of the Regional term (likelihoods of F_g and B_g) $R(A)$ against the boundary term (Relationship between neighborhood pixels) $B(A)$. The regional or the likelihood of the F_g and B_g is given by

$$R(A) = \sum_{p \in P} R_p(A_p) \dots\dots\dots(3)$$

During the minimization of the graph energy formulation in (2) to segment thin objects like blood vessels, the second term (boundary term) in (2) has a tendency to follow short edges known as “the shrinking bias” [14]. This crisis cause a important dilapidation of the routine of the graph cut algorithm on thin elongated structures like the blood vessels.[6]

C. Optic Disc Segmentation

The optic disc segmentation starts by defining the position of the optic disc. This method used the convergence feature of vessels into the optic disc to view its location.

method, the compensation factor approach segments the optic disc using prior local intensity knowledge of the vessels. Fig.3 shows the overview of both the MRF and the compensation factor methods.[1]

D. Optic Disk Location

The double image of vessels segmented is used to find the location of the optic disc. The process iteratively traces toward the centroid of the optic disc. The vessel image is prune via a morphological open method to eliminate thin vessels and keep the main arcade. The centroid of the arcade is calculated using the following formulation:

$$C_x = \sum_{i=1}^K \frac{x_i}{K} \quad C_y = \sum_{i=1}^K \frac{y_i}{K} \dots\dots\dots(4)$$

Given the gray scale intensity of a retinal image, we select 1% of the brightest region. The algorithm detect the brightest region through the most number of pixels to establish the location of the optic disc with respect to the centroid point (right, left, up, or down). The algorithm adjusts the centroid point iteratively until it reaches the vessel convergence point or the center of the main arcade (center of the optic disc) by reducing the distance from one centroid point to next one in the direction of the brightest region, and correcting the central position inside the arcade accordingly.[6]

In contrast to the MRF image reconstruction, we have incorporated the blood vessels into the graph cut formulation by introduce a compensation factor V_{ad} . This feature is consequent using former information of the blood vessel. The power function of the graph cut algorithm usually comprises boundary and regional terms. The boundary term defined is used to assign weights on the edges (n-links) to measure the similarity between neighboring pixels with respect to the pixel proprieties (intensity, texture, and color). Therefore, pixels with similar intensities have a strong connection. The regional term in (3) is derived to define the likelihood of the pixel belonging to the B_g or the F_g by assigning weights on the edges (t-link) between the image pixels and the two terminals B_g and F_g seeds. In order to incorporate the blood vessels into the graph cut formulation, we derived the t-link as follows:

$$S_{link} = \begin{cases} -\ln P_r(I_p \setminus F_{g_{seeds}}) & \text{if } p \neq \text{vessel} \\ -\ln P_r(I_p \setminus F_{g_{seeds}}) + V_{ad} & \text{if } p = \text{vessel} \end{cases} \dots\dots\dots(5)$$

$$T_{link} = \begin{cases} -\ln P_r(I_p \setminus B_{g_{seeds}}) & \text{if } p \neq \text{vessel} \\ -\ln P_r(I_p \setminus B_{g_{seeds}}) & \text{if } p = \text{vessel} \end{cases} \dots\dots\dots(6)$$

Where p is the pixel in the image, F_g seeds is the intensity distribution of the F_g seeds, B_g seeds represents the intensity distribution of the B_g seeds, and V_{ad} is the compensation factor given as

$$V_{ad} = \max_{p \in \text{vessel}} \{-\ln P_r(I_p \setminus B_{g_{seeds}})\} \dots\dots\dots(7)$$

The intensity distribution of the blood vessel pixels in the, region around the optic disc makes them more likely to

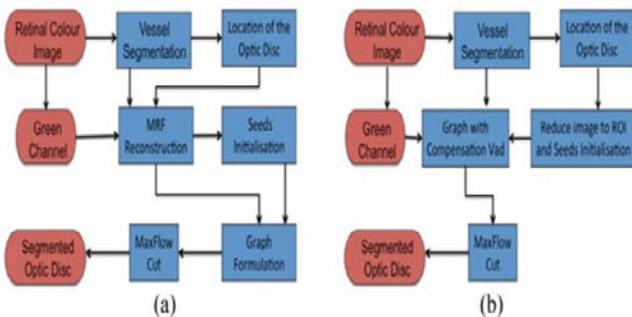


Fig.3. (a) MRF image reconstruction method diagram and (b) compensation factor method diagram

The disc area is then segmented using two dissimilar automatic methods (MRF image reconstruction and compensation factor). Both methods use the union feature of the vessel to discover the location of the disc. The MRF method is useful to eradicate the vessel from the optic disc region. This process is known as image reconstruction and it is performed only on the vessel pixels to avoid the alteration of other structures of the image. The reconstructed image is free of vessels and it is used to segment the optic disc via graph cut. In contrast to MRF

belong to Bg pixels than the Fg (or the optic disk pixels). Therefore, the vessels inside the disk have weak connections with neighboring pixels making them likely to be segmented by the graph cut as Bg.

E. Diseases Classification

Extracted blood vessels and optic disc information's are given to ANN classifier for the detection and categorization of diseases. The function of image categorization scheme is to assign each input to one of sample classes. It is the method of assigning a label to each unknown input image. In this work, the artificial neural network approach namely, Back propagation network (BPNs) is used to classify the images. The back propagation algorithm is used in layered feed-forward ANNs. This means that the artificial neurons are organized in layers and send their signals "forward", and then the errors are propagated backwards. The network receives input by neurons in the input layer, and the output of the system is given by the neurons on an output level.

Here, back propagation algorithm is applied for learning the samples, Tan-sigmoid (tansig) and log-sigmoid (logsig) functions are applied in hidden layer and output layer respectively, Levenberg-Marquardt optimization (trainlm) is used for adjusting the weights as training methodology. For training process, firstly altered features are extracted block by block in one image. When we use a new image for classification, only those elected features are extracted and the trained classifier is used to classify the abnormality in the image.

Diabetic retinopathy is cause by mutually the forms of diabetes i.e. diabetes mellitus and diabetes insipidus. It is a extremely asymptomatic disease in the premature stages and it could lead to lasting vision loss if untreated for long time. The problem here is the patients may not know about it until it reaches advanced stages. Once it reach advanced stages vision loss become unavoidable [4]. As diabetic retinopathy is the third major cause of blindness particularly in India, there is an immediate requirement to develop efficient diagnosis method.

III. RESULT AND DISCUSSION

For the vessel segmentation method, we tested our algorithm on two public datasets, DRIVE and STARE with a total of 60 images. The optic disc segmentation algorithm was tested on DRIVE and DIARETDB1, consisting of 129 images in total. The performances of both methods are tested against number of alternative methods. The DRIVE consists of 40 digital images which were captured from a Canon CR5 non mydriatic 3CCD camera at 45° FOV. The images have a size of 768 × 54 pixels. The dataset includes masks to separate the FOV from the rest of the image. Figs.4 and 5 show the segmented images and the manually labeled images for the DRIVE and the STARE datasets, respectively.

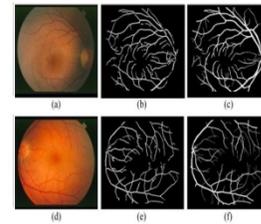


Fig 4 DRIVE dataset

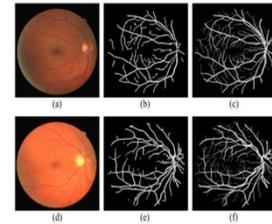


Fig 5 STARE dataset

The entire work done by with the help of MATLAB. Experiment shows that the outcome the scheme is comparable with others when applied on standard data set (images). This is clear in the simulation output shown in Figure 8.1 to figure 8.4

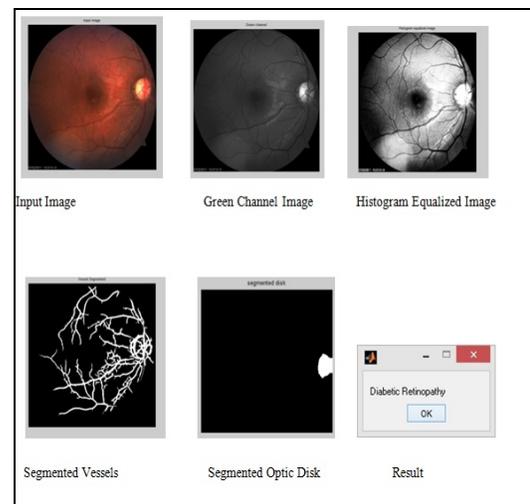


Fig 6: I/O data set

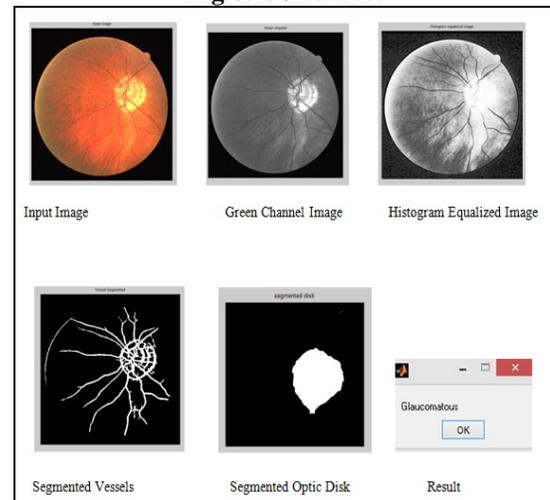


Fig 7 Results of Glaucoma

Fig 6 shows the processing steps for the detection of Diabetic retinopathy. Input image is converted in to green channel image for noise reduction. Blood vessels are segmented by the graph cut technique.MRF and compensation factor method gives the optic disc. Further processing via ANN gives the effected disease. Fig 7 shows the processing steps for the detection of Glaucoma.

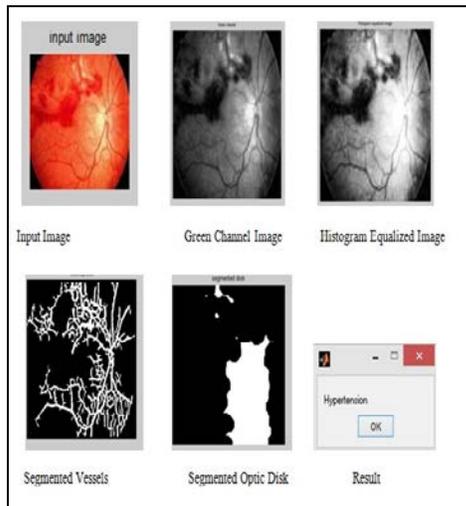
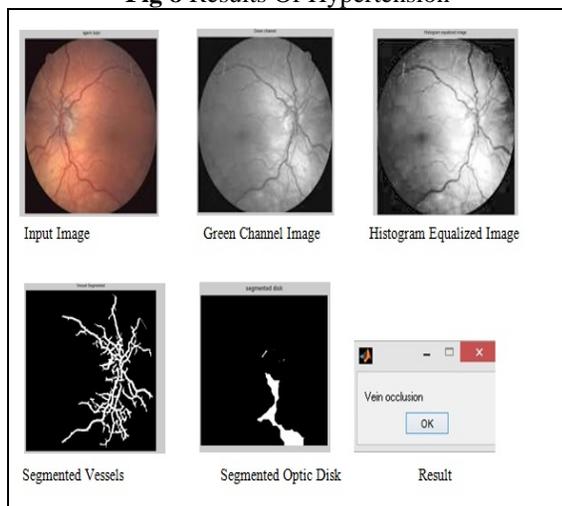
**Fig 8 Results Of Hypertension****Fig 9 Results Of vein occlusion**

Fig 8 shows the processing steps for Hypertension detection. Fig 9 shows the processing steps for vein occlusion detection. The extracted optic disc and blood vessels region is given to an ANN classifier which help in further classification of related diseases like Diabetic Retinopathy, Glaucoma, Hypertension and vein occlusion of the patients.

IV. CONCLUSION

In this paper we have presented blood vessels and optic disk segmentation in retinal images by integrating the mechanism of flux, MRF image reconstruction, and compensation factor into the graph cut method.

In the second stage information's extracted from blood vessels and optic disc are given to an ANN classifier and find whether the image is infected or normal, finally classify the diseases. This proposed methodology can be utilized in hospitals to detect diseases occurring on the eyes by doctors easily. Future scope of this project is to detect many eye diseases thus making mankind to be benefitted in large extent to be free from eye diseases leading to blindness with higher efficiency. From the results and its slotted out puts, clearly identify whole concepts about the whole work.

V. REFERENCES

- [1] D. Welfer, J. Scharcanski, C. Kitamura, M. D. Pizzol, L. Ludwig, and D. Marinho, "Segmentation of the optic disk in color eye fundus images using an adaptive morphological approach," *Comput. Biol. Med.*, vol. 40 no. 1, pp.124–137, 2010
- [2] Soumitra Samanta, Sanjoy Kumar Saha and Bhabatosh Chanda "A Simple and fast Algorithm to Detect the Fovea Region in Fundus Retinal Image", Second International Conference on Emerging Applications of Information Technology, 978-0-7695-4329-1/11\$26.00 ©2011 IEEE DOI10.1109/EAIT.2011.22 206,2011
- [3] C. Sundhar, D. Archana "Automatic Screening of Fundus Images for Detection of Diabetic Retinopathy", *International Journal of Communication and Computer Technologies*, Volume 02 – No.1 Issue: 03 April 2014
- [4] Gowthaman R, "Automatic Identification And Classification of Microaneurysms For Detection Of Diabetic Retinopathy" *IJRET: International Journal of Research in Engineering and Technology* eISSN: 2319-1163 | pISSN: 2321-7308 Volume: 03 Issue: 02, Feb-2014.
- [5] L. Xu and S. Luo, "A novel method for blood vessel detection from Retinal images," *Biomed. Eng. Online*, vol. 9, no. 1, p.14,2010.
- [6] Ana Salazar-Gonzalez, Djibril Kaba, Yongmin Li, and Xiaohui Liu, "Segmentation of the Blood Vessels and Optic Disk in Retinal Images" *IEEE Journal of biomedical and health informatics*, volume.18, No.6 Nov-2014
- [7] Anil K. Jain Michigan State University Jianchang Mao K.M. Mohiuddi ZBMAZmaden Research Center, "Artificial neural networks:A tutorial"
- [8] S. Vicente, V. Kolmogorov, and C. Rother, "Graph cut based Image segmentation with connectivity priors," in *Proc. IEEE Conf.Comput.Vis.Pattern Recogn.*, 2008, vol. 1, pp. 1–8.
- [9] V. Kolmogorov and Y. Boykov, "What metrics can be approximated by geo-cuts, or global optimization of length are and flux," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, 2005, vol.1, pp.564–571.
- [10] Jestin V.K., J.Anitha and D.Jude Hemanth. Genetic Algorithm for Retinal Image Analysis. *IJCA Special Issue on Novel Aspects of Digital Imaging Applications (DIA) (1)*:48–52, 2011.
- [11] G. Salazar-Gonzalez, Y. Li, and X. Liu, "Retinal blood vessel segmentation via graph cut," in *Proc. IEEE 11th Int. Conf. Contr. Autom. Robot.Vis.*, 2010, pp. 225–230.
- [12] Ana Salazar-Gonzalez, Djibril Kaba, Yongmin Li, and Xiaohui Liu, "Segmentation of the Blood Vessels and Optic Disk in Retinal Images" *IEEE Journal of biomedical and health informatics*, volume.18, No.6 Nov-2014
- [13] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Proc. IEEE 8th Int. Conf. Comput. Vis.*, 2001, vol. 1, pp. 105–112.
- [14] S. Vicente, V. Kolmogorov, and C. Rother, "Graph cut based image segmentation with connectivity priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2008, vol. 1, pp. 1–8.

Fuzzy Logic Based Performance Analysis of Various Multiplier Architectures

Vardhana M

Professor, Dept.ECE

Sahyadri College of engineering and management

Abstract— This paper proposes the analysis and evaluation of performance of various multiplier architectures using fuzzy logic techniques. The various parameters of the multiplier, such as timing delay, power dissipation, memory utilization, device utilization, are the inputs to the Fuzzy Inference System and the Performance Index of the Multiplier Architecture will be the output, which evaluates the performance of the Multiplier Architecture. Various Multiplier Architectures like, Karatsuba, Array, Booth, Vedic, and Radix techniques are analyzed for their performance and their performance index are compared. Fuzzy rules are set up according to the advice of experts, taking into consideration timing delay, power consumption, memory utilization and device utilization. Fuzzy Logic based performance analysis of various multiplier architectures proves to be an efficient technique for performance analysis and evaluation. The various multiplier architectures are analyzed and their performance indexes were compared. From the experiment, it was found that the performance index of Vedic multiplier was 0.78 and the performance index for Array multiplier was found to be 0.23. From the comparison of the performance index, it is revealed that Vedic multiplier is the best performer, and Array multiplier is the worst performer, as compared with conventional multiplier architectures.

Keywords— Fuzzy Logic, Fuzzification, Membership Function, Fuzzy Inference System, Multiplier Architecture, Defuzzification.

I. INTRODUCTION

Soft computing is a branch of computing, which can deal with imprecision, uncertainty, partial truth and approximation. The guiding principle of soft computing is to exploit the tolerance for imprecision, uncertainty and approximation to achieve the tractability, the robustness, the low cost solution and the better report with reality. Soft computing consists of Fuzzy logic, Neural Networks, Evolutionary Computation and Chaos Theory. The application of Soft Computing techniques is a good option for many engineering problems, such as modelling, control, classification etc. Fuzzy logic embeds human knowledge into working algorithms; it is well suited in situations involving highly complex systems, whose behaviour is not well understood. In this paper, we propose a new technique of evaluating the performance of various multiplier architecture, using Fuzzy logic, since there exists an uncertainty in decision making about the performance of multiplier architecture. Thus by using the Fuzzy logic, the performance of multiplier architecture can be evaluated, with the help of if-then rules. Multiplication plays a very important role in modern computational systems. Multiplication based

operation such as, multiplies and accumulates [5], convolution, DFT is among some of the frequently used computation. Thus the performance of the multiplier architecture plays a very important role. Multiplication forms the basis of most of the DSP processors. The delay path determines the performance of algorithm. Thus the performance analysis of multiplier architecture forms a very important factor. There are various architecture available for multiplication some of them are, Vedic[1-3], Karatsuba[4], Array[6], Radix[7], Booth, etc. In this paper, we propose a new technique of evaluating the performance of multiplier architecture using Fuzzy logic. Various performances related factors are analysed and performance index is compared. Thus the performance of various multiplier architecture are analysed and evaluated using Fuzzy logic.

II. FUZZY LOGIC AND FUZZY CONCEPT

The fuzzy logic tool was introduced in the year 1965 by Lotfi Zadh, which is an efficient tool for dealing complicated problems where there exists uncertainty. It offers a soft computing partnership which is the important concept of computing with words. It embeds human knowledge into working algorithms. The fuzzy theory provides a mechanism for representing linguistic constructs such as many, low, medium, often few. In general, the fuzzy logic provides an inference structure that enables appropriate human reasoning capabilities. On the contrary, the traditional binary set theory describes crisp events, that either do or do not occur. It uses probability theory to explain an which occurs by event measuring the chance with which a given event is expected to occur. The fuzzy logic theory is based upon the notion of relative graded membership which are the functions of cognitive processes. The utility of fuzzy sets lie in their ability to model uncertainty or ambiguous data. Fuzzy logic can be used for various other applications which may include the performance analysis of students, teachers [8-10].

III. IMPLEMENTATION OF FUZZY INFERENCE SYSTEM

Fuzzification involves conversion of crisp values into grade of membership. Fuzzy sets can represent the degree to which a quality is possessed. Fuzzy sets have the values in the range $[0, 1]$. The method of conversion of crisp values into fuzzy linguistic values is called as fuzzification. Fuzzy rules can be set with the help of simple if-then rules which can be utilized

for decision making in the situation of uncertainty. In this paper the performance deciding factors like delay, device utilization and power are the crisp values which is converted into linguistic variables. The linguistic variables may be good, very good, excellent etc. Depending upon the if-then set of rules fuzzy inference system takes the decision about the performance of the multiplier architecture.

Fuzzy inference systems (FISs) are also known as fuzzy rule-based systems, fuzzy model, fuzzy expert system and fuzzy associative memory which forms a major unit of a fuzzy logic system. The decision-making is an important part of the entire system. The FIS formulates suitable rules and based upon the rules the decision is made. The basic FIS can take either fuzzy inputs or crisp inputs, and produces output which are almost always fuzzy sets. Fig.1. shows the FIS editor using fuzzy tool. Input to the Fuzzy Inference System is the performance determining parameters such as delay, power, memory, device utilization. These parameters are analyzed according to the if-then rules to give the performance index in terms of fuzzy linguistic variables. These linguistic variables are converted to crisp values to get the performance index. Fig.2. shows the output membership function for performance index. Fig.3. shows the Fuzzy inference system for device utilization which accepts the inputs of number of slices, LUTs etc. Fig.4. shows the Fuzzy Rule Viewer for the if-then rules. Fig.5. shows the Fuzzy rule viewer for the device utilization. Fig.6. and Fig .7. Shows the 3 dimensional variation of the parameters.

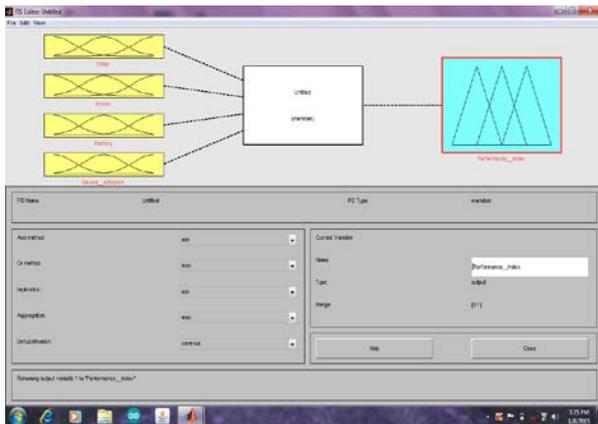


Fig.1. FIS Editor

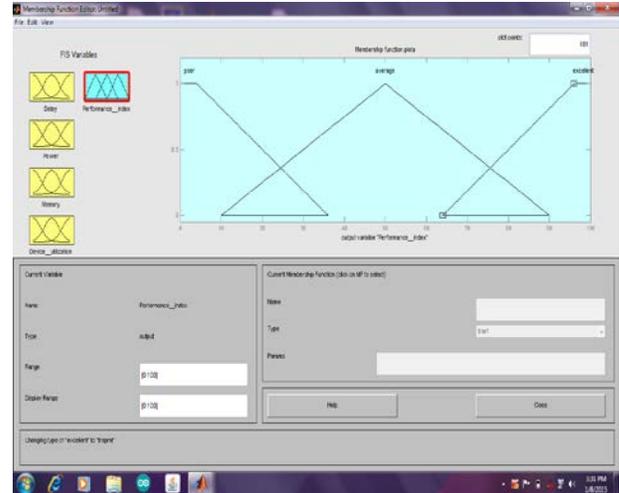


Fig.2. Rule Viewer

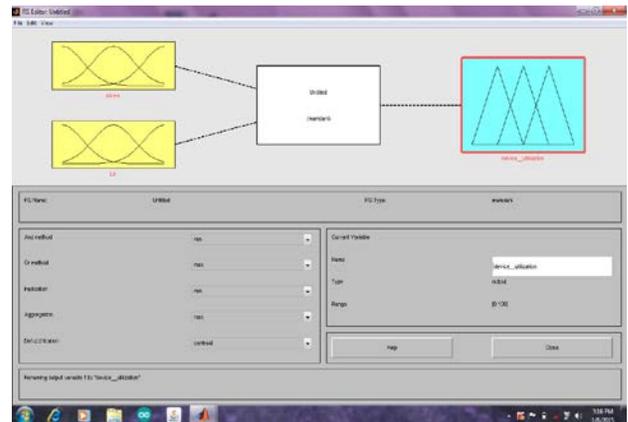


Fig. 3. Device utilization

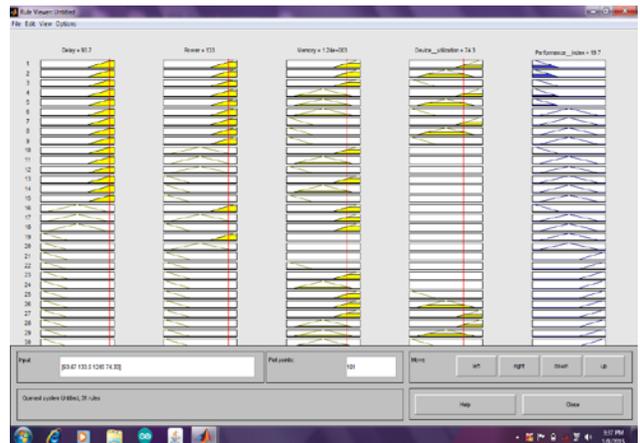


Fig. 4. Fuzzy Rule Viewer

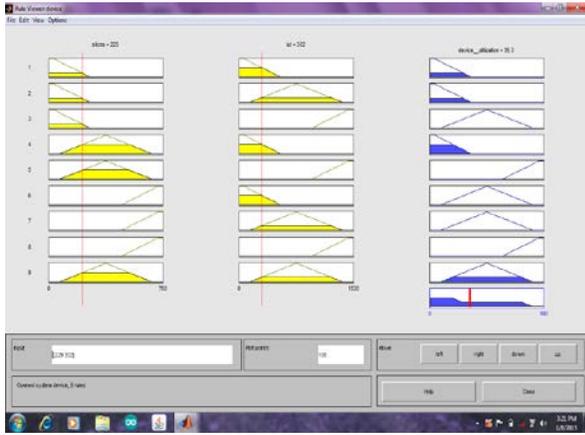


Fig. 5. Device Utilization

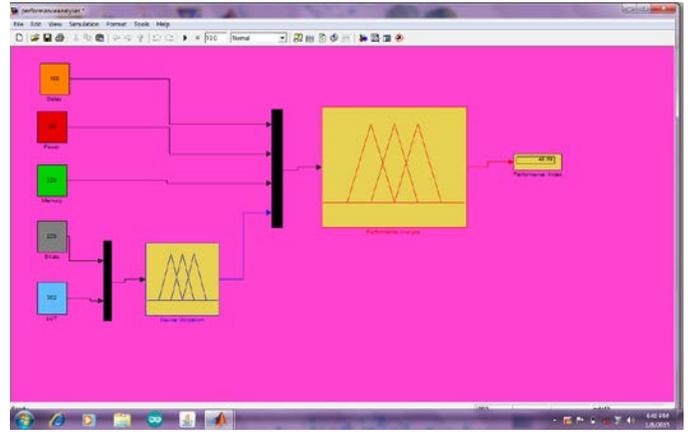


Fig.8. Modeling Using Simulink

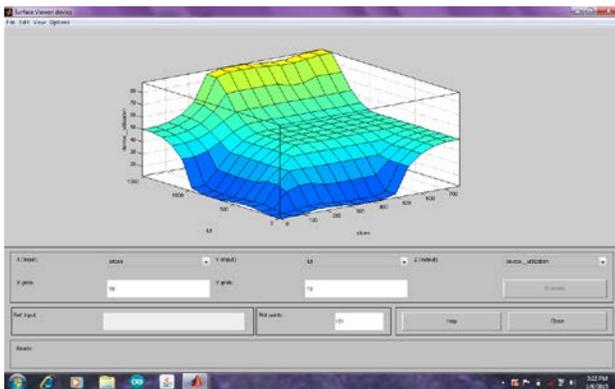


Fig.6.Surface View for Device Utilization

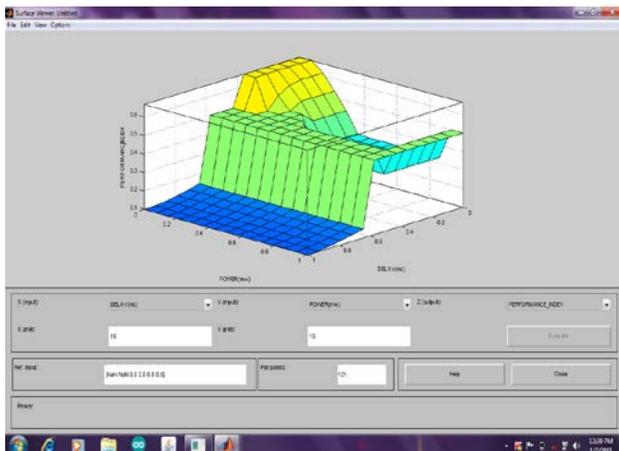


Fig.7. Surface Viewer

Thus developed Fuzzy system is invoked using the simulink and the performance evaluation system is modeled. The input to the system is various performance related parameters and the parameters are evaluated and performance index is displayed. Figure.8. shows the modeling of the system using Simulink

IV. RESULTS

The efficiency of any system depends upon the performance of ALU unit in it, which in turn depends on the multiplier architecture. Hence the efficiency of the multiplier architecture is one of the major area of concern. In this paper Fuzzy Inference System is developed for the accurate analysis of the performance of various multiplier architecture. The input to the system involves speed, delay, power etc, which are then analysed to give the performance index. Finally the system modeling is done using Simulink.

The various Multiplier architecture are analyzed for their performance and it is found that, the performance index of Vedic multiplier was 0.78 and the performance index for Array multiplier was found to be 0.23. From the comparison of the performance index, it is revealed that Vedic multiplier is the best performer, and Array multiplier is the worst performer, as compared with conventional multiplier architectures.

ACKNOWLEDGMENT

The author(s) thank student and faculty of Sahyadri College of Engineering and Management.

REFERENCES

- [1] M. Poornima, ShivarajKumarPatil, Shivukumar, K.P. Sridhar, and H. Sanjay, "Implementation of Multiplier using Vedic Algorithm," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol 2, no. 6, May 2013.
- [2] G. Ganesh Kumar, and V. Charishma, "Design of High Speed Vedic Multiplier using Vedic Mathematics Techniques," International Journal of Scientific and Research Publications, vol2, no. 3, March 2012.
- [3] Jagadguru Swami Sri Bharath Krsna Titartji, Vedic Mathematics or Sixteen Simple Sutras from the Vedas, Motilal Banarsidas, Varanasi, 1986.
- [4] Laszlo Babai, Algorithms-CMSC-37000 Divide and Conquer: The Karatsuba-Ofmanalgorithm, <http://www.cyclopaedia.info/wiki/Karatsuba-Multiplication>.
- [5] Anvesh Kumar, and Ashish Raman, "Low Power ALU Design by Ancient Mathematics," IEEE Transactions on Computing, 2010.
- [6] Z. Huang, and M.D.Ercegovac, "High-performance low-power left-to right array multiplier design," IEEE Transactions on Computing, vol.54, no.3, pp.272-283, March 2005.

- [7] S. Jagadeesh, and S. Venkata Chary, "Design of Parallel Multiplier-Accumulator Based on Radix-4 Modified Booth Algorithm with SPST," International Journal of Engineering Research And Applications (IJERA), vol.2, no. 5, pp.425-431, September-October, 2012.
- [8] Sirigiri Pavani, P.V.S.S Gangadhar, and Kajal Kiran Gulhare, "Evaluation of Teacher's Performance using Fuzzy Logic Techniques", International Journal of Computer Trends and Technology- volume3Issue2- 2012
- [9] Hota H.S., Sirigiri Pavani, P.V.S.S. Gangadhar, "Evaluating Teachers Ranking Using Fuzzy AHP Technique", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013
- [10] E. Sakthivel, K. Senthamarai Kannan and S. Arumugam, "Optimized Evaluation of Students Performances Using Fuzzy Logic", International Journal of Scientific & Engineering Research, Volume 4, Issue 9, September-2013, ISSN 2229-5518.

Renovation CoReVDO[®]

Methodology of Collaborative Requirements Validation in Distributed Organizations

Sourour Maalem

Department of Computer Science

ENS Constantine

E-mail: soumaalem@yahoo.fr

Abstract—The work we present in this paper is devoted to an important and delicate activity of the Requirements Engineering (RE) which is the validation applied to collaborative environments, it comes as complementary to the methodology CoReVDO[1] where we considered as a prospect the management of linguistic properties of requirements, such as the key elements to classify functional requirements with respect to the desired qualities in a specification. The Renovation of the CoReVDO[®] methodology (Collaborative Requirements Validation in Distributed Organizations) provides a process divided into three phases: organizational phase, distributed requirements verification phase and collaborative validation of needs phase. This renovation involves the separation between the system requirements and user needs; between requirements validation and requirements verification between the functional and nonfunctional requirements, and finally between the functional criteria of semantic aspect and functional criteria of non- semantic aspect.

Keywords —Requirement/Need, Requirements Engineering, RE Methodology, Requirement Validation, Requirement Verification.

I. INTRODUCTION

The Requirements Engineering (RE) discipline is one of the most important phases in the traditional cycle of software development, as it helps to determine the purpose of a system properly. Its purpose is to identify and document the stakeholders and their needs in a form allowing their analysis, communications, and implementation. The basic elements of this discipline are the needs and requirements. The requirements represent the system vision from a designer perspective (technical perspective), the need is the system vision from a user perspective. A requirement is a need that is either technically satisfiable or its solution can be implemented. Requirements formalize the expression of needs and commitments of stakeholders E. Didier et al [6].

The increasing complexity of systems and the interdisciplines necessary to conduct them push organizations to use teams, because it seems like a way to improve the quality and time of realization of industrial products. Increasingly, cooperation is the subject of much research and the number of conferences devoted to it demonstrates the importance of this subject. However, it is not a recent practice because people always cooperate. However, cooperation as it reappears today is not just a team, it means structured and organized to produce better results. The quality assurance of user needs and system requirements is ensured by the steps Verification and Validation (VV).

Activity Verification and Validation in RE helps to improve the quality of a requirements specification document and, therefore, the success of a project. The idea of this work focuses on how to validate the requirements by the masters of works and / or building owners who are geographically distributed. For this, a collaborative validation will take place online. It is based on the different views of stakeholders on the quality of specified requirements and negotiated in a virtual meeting to generate a prototype that will be distributed in order to be accepted, rejected or modified.

To go through our ideas, in addition to the lack of methodologies for validating requirements, particularly in the cooperative context, we face other kinds of problems. Those related to the cooperative platform (open universe, distribution, heterogeneity, conflict situations, jet lag, etc.). Other problems are directly related to process validation requirements (granularity act or activity in Phase What is the validation result What to validate , what check what types of processes are appropriate VV How can all agree ? ? stakeholders How to verify and validate , etc. .) .

We try to answer these questions in the methodology we propose, called CoReVDO[®] (Collaborative Requirements Validation in Distributed Organizations). CoReVDO[®] is developed based on an approach distinguishing between system requirements and user needs between the requirements validation and requirements verification between functional and nonfunctional requirements between the functional criteria of semantic aspect and functional criteria of non- semantic aspect, and multi views following a distributed and collaborative verification and validation process.

The paper is organized as follows: Section 2 gives an overview of some similar work. Section 3 describes the methodology for the validation of proposed CoReVDO[®] requirements. Section 4 concludes and gives some perspectives.

II. RELATED WORK ON METHODOLOGIES VALIDATION REQUIREMENTS

Most of the existing requirements validation methods (VE) aim only to identify and collect missing and conflicting requirements. Due to time and other considerations validation is done informally, on an ad - hoc basis or simply as a

peer_review exam . Most of them only provide guidelines on how developers and customers should review the requirements specification to find inconsistencies and errors. A. Katasonov [11] outlines the requirements validation as a list of "best practices" based on the use of a large number of heterogeneous techniques. L. He et al [12] discovered that the validation requirements are often not sufficiently covered not only in the practical world, but even in the academic world. The literature mentions two levels of work: an external validation that is the proper sense of the word, with the user, the client and / or other stakeholders, based on techniques such as interactions, prototyping, animation or simulations. Another internal audit is usually this activity is performed by the IE team is most often based on non- friendly techniques such as mathematical formulation , formal logic, natural language processing using tools complex . This level still needs to be supplemented by a second iteration of processing so that it can be approved by the building owner, if credibility is questionable.

In the current work there is a mix, it works on several levels with internal and external modes (formal, semi-formal, informal) using manual methods and / or automated, see Table 1.

In terms of the validation process in IE, R. Cavada et al [7] propose a validation process subdivided into fragmentation and categorization, formalization and validation, they provide a tool that analyzes natural language requirements for classified and structures. Then formalize requirements subsets of UML enriched with static and temporal constraints for which it was defined a formal semantics. Finally, the tool allows to apply techniques of "model checking" specialized for the validation of formal requirements. In [8] , P. Scandurra et al automatically transform use cases into executable ASM specifications and validate functional requirements through simulation based on the ASM generator using ASMETA of analysis tools.

In [2] , R. Cavada et al combine formal verification engine and MS Word ® editor in a single environment consisting . Based on a plug-in, that uses the Python language along with a MS Word ® Add-In document. The user can jump between textual requirements in MS Word ® editor in the corresponding formal requirements model. A. Chiappini et al [9] after a validation process that progress from an informal analysis formalization to reach a formal validation. They develop a set of tools that support various phases and apply it to a real subset of a specification of industry.

All these works begin with informal requirement, turn them into formal or semi formal models to check certain constraints. Which consumes more time , effort and cost compared to what is expected for the IE phase and consequently the total duration of the project and the need for experts in formal methods analysts , which is not always the case (analysis, planning , negotiation ...) .In contrast to the approaches mentioned above , the QualiCES method where L Muriana M. et al [10] propose a wraparound approach begins by

identifying the needs of quality followed by an analysis of the documents and then applying a " Checklist " accompanied by the application of consistent metrics and ends with the analysis of results . Without making call to formal method.

WORK	PROCESS	VALIDATION LEVEL		VALIDATION TECHNIQUE	TOOLS USED	ANIMATOR VV
		INTERNAL	EXTERNAL			
R.CAVADA ET AL. [7]	1. FRAGMENTATION AND CATEGORIZATION 2. FORMALISATION 3. VALIDATION	MODEL CHECKING UML CNL	MS WORD	INFORMAL SEMI-FORMAL FORMAL	EURAILCHECK	MULTI USER
P.SCANDURRA ET AL. [8]	/	USES CASE SCENARIOS ASM	SIMULATION	SEMI FORMAL FORMAL	ASMETA	MULTIUSER
R. CAVADA ET AL. [2]	/	PYTHON GTK C#	MS WORD® ADD-IN	GRAPHIC FORMAL	OTHELLOPLAY	MULTIUSER
A.CHIAPPINI ET AL. [9]	1. INFORMAL ANALYSIS 2. FORMALIZATION 3. FORMAL VALIDATION	UML2 PSL LTL	ECLIPSE	FORMAL	RSA REQUISITEPRO	MULTI USER
L M MURIANA ET AL. [10]	1. IDENTIFICATION OF QUALITY NEEDS 2. ANALYSIS NECESSARY DOCUMENTS 3. APPLICATION OF THE CHECKLIST 4. APPLICATION OF CONSISTENT METRIC 5. ANALYSIS OF RESULTS	METRIC USE CASE	CHECKLIST PROTOTYPE INTERFACE	INFORMAL SEMI FORMAL	QUALICES	MONO USER
S.MAALEM, N. ZAROUB [1]	1. PHASE ORGANIZATION 2. PHASE DISTRIBUTION VERIFICATION 3. COLLABORATIVE VALIDATION PHASE	QFD, CHECK-LIST, GROUPWARE	PROTOTYPE	INFORMAL SEMI FORMAL	WHITE BOARD	DISTRIBUTED HETEROGENEOUS GROUP

TABLE I. COMPARAISON AND CLASSIFICATION

To our knowledge the only link between the two domains (Cooperation & engineering requirements) and suggests approaches for validating requirements is CREWS (Cooperative Requirements Engineering With Scenarios) [23]. This approach aims to develop, evaluate and demonstrate methods and tools for the elicitation and validation of requirements based on cooperative scenarios. Through two approaches CREWS guide the different uses of scenarios

during the validation requirements. The first approach A. Neil et al [5] provides a method and software – Crews-Savre tool to generate scenarios and provide a review of such Walkthrough. The second approach to the validation process requirements in P.Heymans [13] is carried out through the animation of scenarios from a formal specification. One of the strengths of CREWS approaches is that they were designed to complement each other. However, they have some weaknesses, CREWS has a much more collaborative requirements elicitation method, but far from being a method of validation for cooperative scenarios representations provide a single point of view. According to Sommerville et al [16], a single perspective (the perspective of system requirements) cannot find a set of plain requirements. Views reflect the skills, objectives and roles of each participant, so we believe that the use of a multi-viewpoints approach to validate is almost inevitable in a collaborative organization. Even the appearance of cooperation as a principle in CREWS has not been sufficiently supported, since the distribution of stakeholders, their heterogeneity and the maintenance agreement between the various stakeholders in the event of negotiation and / or cooperation remained neglected. Moreover, the overlap between the elicitation phase and the specification in CREWS makes it very difficult to document due to move to the design phase. CoReVDO in [1], we presented a collaborative approach to the validation process requirements based on the concepts of multi skills and perspectives. The approach is divided into three phases : organization, distributed collaborative verification and validation process without the specific system requirements and user needs between the requirements validation and verification requirements between functional and non functional requirements and between the functional criteria of semantic aspect and functional criteria of non-semantic aspect, which will be the subject of this article.

III. METHODOLOGY CoReVDO®

The involvement of three research areas: Cooperation, RE and Validation /Verification (VV) in this paper lead to three classes of problems:

- Problems with the definition of validation in the literature.
- Problems related directly to the process of validation requirements (granularity activity or stage, what is the validation result, What to validate, what check what types of processes are appropriate VV, How to deal all stakeholders?? How to verify and validate, Etc.).
- Problems related to the cooperative platform (open universe, distribution, heterogeneity, conflict situations, jet lag , etc.).

1. Solve the problem of the ambiguity of the term validation

F.Fabbrini et al [17] proposed four types of language requirement properties (syntactic, structural, semantic and

pragmatic) as key elements to classify functional requirements (see Figure 1). The semantic property must be validated through human interaction, but non-semantic (syntactic, structural and pragmatic) can be checked manually or automatically.

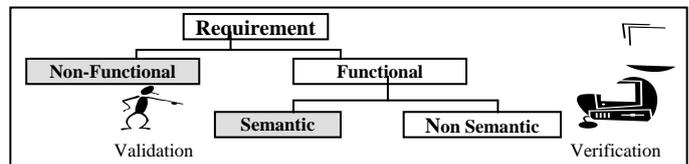


Fig. 1. Requirements taxonomy [17]

Katasonov et al [11] presented a link between the quality properties of a specification (IEEE 830) and requirements VV process as follows.

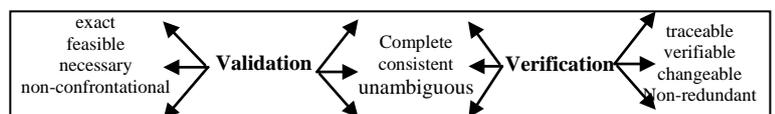


Fig. 2. Link properties compared to VV [11]

By logical sequence that comes from these two explanations Fabbrini Katasonov and we are redefining the audit requirements as "syntactic examination, structural and pragmatic functional requirements not semantic orientation to ensure the quality criteria for traceability, the No redundancy, modifiability and verifiability, internally with less means more compounds according to the ability of the prime contractor.

In distinction to the validation requirements "which is responsible for characterizing the semantics and non-functional aspect, to certify the completeness, the consistency, the No ambiguity Not the conflict, the accuracy, the Feasibility and Necessity" requirements where its external validation is not final until approval of the master developer of the need to use more simple means, friendly and pleasant.

2. Resolve the problems related with the validation in RE process.

Our view of the validation requirements is broader than that found in the literature because it addresses the validation of such a process of verification of system requirements by experts followed by a bargaining agreement between the requirements to experts which will pass the validation requirements from customers (Figure 3).

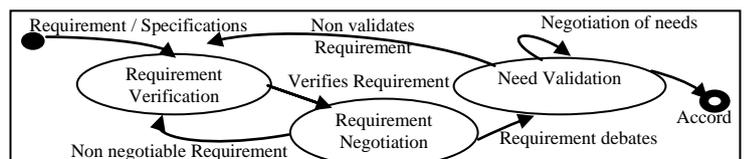


Fig. 3. Validation requirements process of CoReVDO® methodology

The validation process requirements CoReVDO[®] methodology is progressive in that each activity is affected to a reasoned one or more stakeholders. It operates on the use case diagram in Figure 4:

The selection of participants is based on certain criteria which differ according to the phases. We will focus in particular have different perspectives and make use of multiple and complementary skills. Generally, it is best to choose the participants outside the development team. They will be more objective because they are less involved in the project. It is then necessary to assign a role to each and it is clear that a participant may possibly hold several roles.

communication (Especially Internet tools and extranet), at this level it will use the Groupware tools that can support the engineering process (point means of communication view), the approach aims to develop. Other emergent properties of cooperative work:

• **Identification:** Identify the uniqueness and importance and / or stability of each requirement, which is usually, presented by an index *i*, this number is a unique number requirement, but in a distributed environment, we must indicate the source of this information. So identifying uses two indexes [*i*, *j*] where *i*: presents a number of requirements with respect to all requirements and *j* identify the person who issued this requirement, which is supposed to come later in the validation and verification.

In addition to the properties that ensure the quality requirements and deliverables (IEEE 830), he reveals other properties and new definitions related to the cooperative platform such as:

• **Complementarity:** in a context where the expertise is complemented by the involvement of other experts that a requirement is incomplete, ambiguous to the other information emanating from other areas or expert response. At a level of iteration, the requirement is not valid until the appearance of other details that justify the need for collaboration in order to complete the sense of the entities involved.

$$\text{Complementarity} = \sum_{i=1}^N \text{Total number of lines QFDs} - \text{the number of lines after QFD Fusion. / } N \text{ is the number of organizations}$$

• **Global Coherence:** They must not contradict other requirements established after the global integration requirements.

$$\text{Global Coherence} = \text{Number of conflicts resolved} / \text{Total conflicts}$$

• **The Similarity:** Identifying similar requirements (which is the same semantics) or overlap in the same specification or between projects. If the size of the requirements set *E* is important is that after the merger is highly complementary skills are otherwise similar. So cooperative work added nothing, here we have two cases: either the choice of collaborating organizations is not the best if not the point of views are very close

$$\text{Similarity} = \forall x, \exists y : y \cup E = x \cup E$$

E: sets of requirements **y:** necessary requirement

4. Presentation of the different phases of the CoReVDO[®] methodology.

The idea of this work focuses on the steps that will take a set of documents (containing requirements) to be validated by

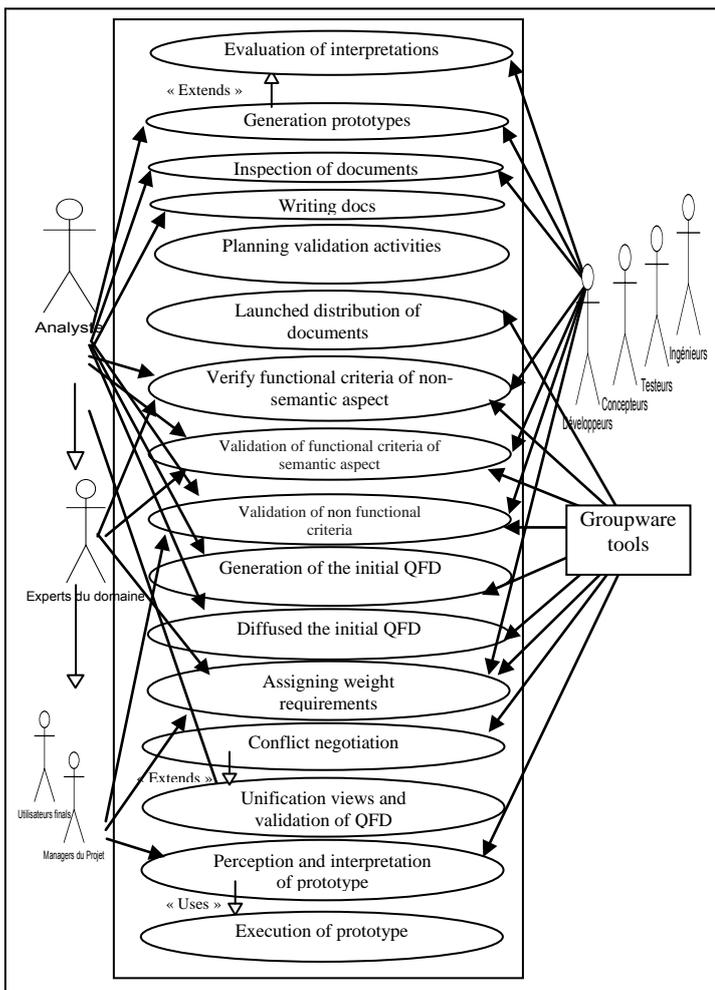


Fig. 4. Usecase diagram of the CoReVDO[®] methodology

3. Resolve problems the cooperative platform

The focus of our approach is multi-view, to establish an IE approach of cooperative information systems. It aims to establish a process to approach a virtual organization consisting of several geographically dispersed organizations where each organization is supported by IS. It will use the resources provided by the new information technology and

the end user, some and / or all of the stakeholders that are distributed geographically.

The new objectives of CoReVDO[®] such as the separation between the system requirements and user needs, requirements validation and verification requirements, functional and non-functional requirements and functional requirements of semantic aspect and criteria functional semantics are not insured by the definition of validation and verification requirements see requirements section (III. 1), include the verification and negotiation as sub-steps in the validation activity see section (III.2) and tested in more quality properties (IEE830) complementarity, overall consistency and similarity see section (III.3) to assess the requirements / needs.

The dynamic operation of CoReVDO[®] is symbolized by the activity diagram of figure 5, the whole process goes through several stages involving various stakeholders. For this, we decompose the validation activity in three phases in order to adapt to a distributed cooperative environment are:

Organizational phase: where goes the development of collaborative validation work through three stages (see Figure 5): (i) planning stage of collaborative work, (ii) stage of qualification requirements, and (iii) step inspection of documents and compliance with the standard.

Phase distributed verification requirements: at each organization, taking place checks the party concerned. Here the work is going through four activities: (i) Activity Verification functional criteria of non-semantic aspect to judge Traceability, Not the redundancy, the modifiability, the Verifiability and Compliance with standard requirements. (ii) Activity Validation functional criteria of semantic aspect to evaluate the completeness, the consistency, the No ambiguity Not the confliction, the accuracy, the feasibility and necessity of the requirements. (iii) Activity Validation of non- functional requirements and (iv) the activity generation of QFD in which performs the allocation of weight in each perspective on the part of the work involved in any organization part.

Collaborative validation phase needs: This phase involves the stakeholder user is guided by the director. The method is multi- perspective and is as follows : (i) Assessment of phase conflicts from different QFD, where a discussion is carried out during the negotiation phase of the conflict (i1), followed by a discussion of agreement in the phase of unification of view (i2) after complete satisfaction of various collaborators, and finally a projection of all the information contained in the QFD to an executable software prototype by the building owner, presented in (ii) the generation phase prototype.

1. Organizational phase

The main objectives of this phase are the development of the working environment, the assignment of roles of stakeholders, collection of documents, preparation of appropriate artifacts, rapid inspection of documents to their dissemination. It is divided itself into three processes: Planning, Inspection and Qualification.

a) planning

This process determines the appointment of meetings through sending mail or using shared by all members of the

engineering team (members of the virtual organization) agenda. Coordination and cooperation of each partner are overseen by the analyst assigns roles

b) Qualification

It is in this process that begins the work of the analyst who performed: (1) The collection of documents produced understandable previous activities by all stakeholders. For the language problem, we suggest documents written in English with the ability to use glossaries and machine translators. (2) The preparation of appropriate artifacts for monitoring the process of validation, so a sheet for each requirement, the model is used FLY, a checklist to identify questions to answer and objectives [1], a rubric to help score satisfied unmet criteria for each requirement separately.

c) Inspection

The IE team meets to review the documents online. Each working according to his expertise and competence; meeting then ends by formulating a checklist that will guide the rest of the process.

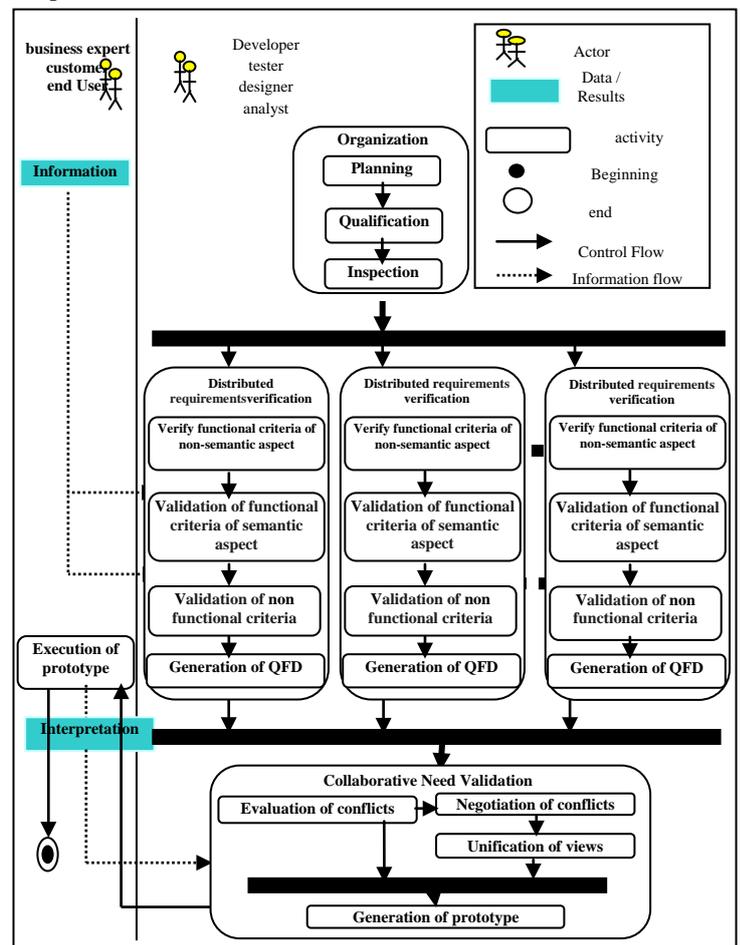


Fig. 5. Activity diagram of the operation of the CoReVDO[®] methodology

2. Phase distributed requirements verification

The purpose of this phase is to ensure the ability of requirements to their specifications in relation to all members of the virtual organization.

This phase is performed by four processes. (i) Process of Verification functional criteria with non- semantic aspect, (ii) Process of Validation functional criteria with semantic aspect (iii) Process Validation of non- functional requirements, and (iv) Process of generation QFD in which performs the allocation of weight in each perspective on the part of the work involved in any organization apart.

a) Process of verification functional criteria with non-semantic aspect:

These criteria are summarized in the review syntactic, structural and pragmatic functional requirements (Figure 6) summarize the properties (Traceable, Non- Redundant, Editable, verifiable and Compliance Standard)

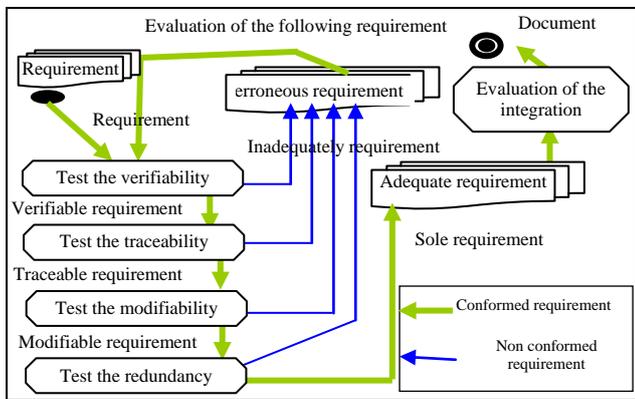


Fig. 6. Verification process of some non-semantic criteria

b) Process of validation functional criteria with semantic aspect:

These criteria are summarized in the semantic examination of functional requirements (Figure 7) summarizes the properties (Completeness, Consistency, Not ambiguity Accuracy, Feasibility and Necessity).

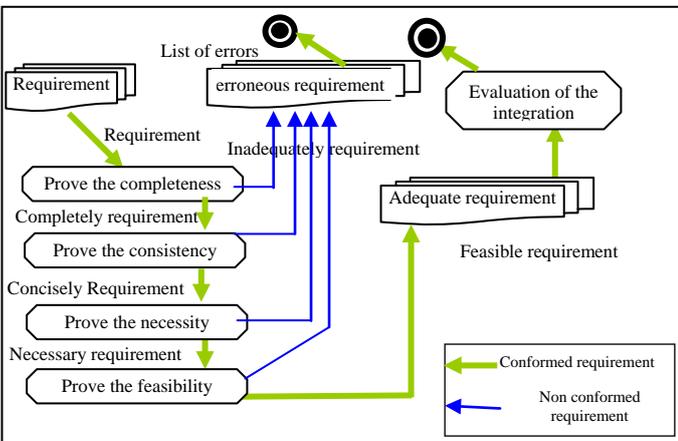


Fig. 7. Validation process of some semantic criteria

c) Process Validation of non-functional requirements:

This evaluation covers three main classes according to the classification requirements of G.Kotonya et al [14]: (1) compliance requirements produced such as usability, efficiency (performance and space), reusability and portability (2) compliance with organizational requirements such as delivery, implementation and standardization (3) compliance with external requirements such as interoperability, legislation (Safety, etc.). This process is incremental, continuous, and progressive ends by filling the evaluation grid for each non-functional requirement, the construction of a technical document and a list of technical errors (Figure 8).

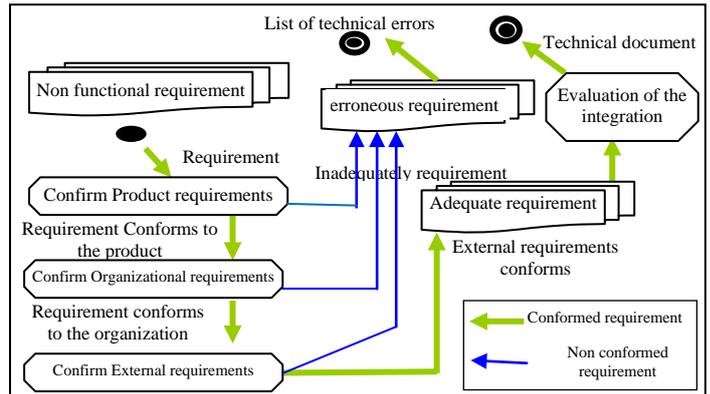


Fig. 8. Validation of non-functional requirements

d) The process of generating QFD

After evaluating all the requirements, they should be with their project specifications in a QFD matrix. The quality function deployment (QFD) is a technique used to analyze customer needs and link them to actions that meet. It uses a matrix linking the requirements and specifications to meet them. The motivation behind the choice of QFD is a technique that is used to analyze customer needs and link them to actions that meet, it is possible to evaluate at the same time the requirements and specification which avoids errors associated with the transition from one phase to another. This process of realization of the QFD is performed automatically by means of a Microsoft Excel spreadsheet or by specialized software (SQFD Client / server version or QFD capture, etc.).

3. Phase collaborative needs validation

After the generation of QFDs a virtual panel is designed with a synchronous type of groupware, and invisible closed allowing unify QFDs groups.

This phase involves the client guided by the contractor and takes place in the following stages: Evaluation of conflict, conflict negotiation, unified views and an executable software prototype for the client.

a) Evaluation of conflict

The input of this phase is all QFDs generated. The commonalities between the expertise of organizations create points of conflict where state values conflicting weight affected by several sources. Different participants (origins) can assign different weight values according to their views for the same torque (requirement specification). This requires a negotiation between these origins.

b) *Negotiation of conflict*

Convince all stakeholders so that they understand the basic requirements of their point of view. The negotiation phase can appear as a separate approaches in some RI activity Kotonya et al [15] and I. Sommerville [16]. Our approach considers negotiation as an inevitable and essential task in the process of validation requirements to find a compromise between the different stakeholders. Information extracted from each local information systems can generate semantic conflicts (conflicts domain definition conflict in interpretation, etc.). Thus, we seek to reach a consensus to negotiate the weight given to the requirements specifications. The choice of a qualitative correlation is to identify the most appropriate link.

c) *Unification of views*

Reach an agreement on a coherent set of requirement that meet the possible stakeholders, is the main objective of the third phase.

Three alternatives for obtaining correlations in QFD have been documented in the literature: (1) application of individual responses and the average results using a moderating factor A. Stylianou et al [18], (2) using a multi-criteria analysis H.In et al [19] P. Chuang [20], and (3) the parties must negotiate their different points of view until a consensus is reached V. Bouchereau et al [21].

Here the metric is applied complementarity between requirements and subsequently evaluating their overall coherence and check if there is no similarity in the specification documents. The latter approach is considered beneficial for the formation of teams, increased participation in product development, obtaining a general consensus on "what to do", and the preservation of momentum when the group rates.

d) *Generation of prototype*

The validation is done by the client or its representatives upon verification. It aims to ensure a match between what the analyst wrote and what the customer has in his head. To do so, the models developed must be communicated to users in order to be accepted, modified or rejected as a prototype.

IE group will go to the automation of an executable rapid prototype to simulate the future system, and this by transforming the QFD (set of specifications verified and negotiated) a set of software interfaces executable line. This semantic difficulty requirements could be described as "articulatory" in reference to the theory of the action of

Norman [22]. This one combines the performance of a task at a distance course. Articulatory distances reflect the difficulties to adapt for the user to commands, and to interpret the state of the system from the state of the interface.

IV. CONCLUSION

The validation exercise is long, expensive and sometimes brings bad news: anomalies, non-compliance and ultimately delay the project, or its outright cancellation. We made our contribution to the process of validation requirements in a distributed environment by providing a methodology for validation of requirements in distributed collaborative organizations named in CoReVDO[®] is what we have:

Dominates the problem of ambiguity of terms VV choosing one definition and distinguishes validation check, specifying who does what, when, where, how and by what means and a process (audit, negotiation, validation), modeled by functional digraphs.

The proposed methodology in CoReVDO[®] approach includes a set of activities required to verify, negotiate, validate and develop a system providing an economic and efficient solution to a client's needs while satisfying all stakeholders. CoReVDO[®] is based on a continuous process, progressive, collaborative, distributed and multi-viewpoints.

We will consider opportunities as the application of this methodology on a case study in order to learn more. Thus, the assumption of phase separation which satisfy the requirements of their individual groupings (requirements document). Finally use metrics and measures to strengthen the methodology during the V V.

REFERENCES

- [1] Sourour. Maalem, N. Zarour, "A methodology of Collaborative Requirements Validation in a cooperative environment" Programming and Systems (ISPS), 2011
- [2] R. Cavada, A. Cimatti, A. Micheli, M. Roveri, A. Susi, S. Tonetta, "OthelloPlay: a plug-in based tool for requirement formalization and validation" May 2011 TOPI '11: Proceedings of the 1st Workshop on Developing Tools as Plug-ins ACM
- [3] G. Gabrysiak, H. Giese, A. Seibel, "Deriving Behavior of Multi-User Processes From Interactive Requirements Validation"; September 2010 ASE '10: Proceedings of the IEEE/ACM international conference on Automated software engineering
- [4] E. Andrade, P. Maciel, G. Callou, B. Nogueira, C. Araújo, "Mapping UML sequence diagram to time petri net for requirement validation of embedded real-time systems with energy constraints" March 2009 SAC '09: Proceedings of the 2009 ACM symposium on Applied Computing ACM Request Permissions.
- [5] A. Neil, M. Maiden, S. Minocha, K. Manning, M. Ryan, "CREWS-SAVRE: Systematic Scenario Generation and Use". ICRE 1998: 148-155.
- [6] E. Didier, "La méthode B et l'ingénierie système. Réponse à un appel d'offre.", Technical report, IUT-Nantes, Université de Nantes, <http://www.iutnantes.univ-nantes.fr/habrias/dessGledn/>, 2002.
- [7] R. Cavada, A. Cimatti, A. Mariotti, C. Mattarei, A. Micheli, S. Mover, M. Pensallorto, M. Roveri, A. Susi, S. Tonetta, "Supporting Requirements Validation: The EuRailCheck Tool" ASE '09:

- Proceedings of the 2009 IEEE/ACM International Conference on Automated Software Engineering, IEEE Computer Society
- [8] P. Scandurra, A. Arnoldi, T. Yue, M. Dolci, "Functional Requirements Validation by transforming Use Case Models into Abstract State Machines". SAC '12 Proceedings of the 27th Annual ACM Symposium on Applied Computing
 - [9] A. Chiappini, A. Cimatti, L. Macchi, O. Rebollo, M. Roveri, A. Susi, S. Tonetta, B. Vittorini, " Formalization and Validation of a subset of the European Train Control System", ICSE '10 Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - Volume 2 Pages 109-118
 - [10] L. M. Muriana, C. Maciel, F. F. Mendes, " QualiCES, A Method for Verifying the Consistency Among Documents of the Requirement Engineering Phase", (2012) SIGDOC '12 Proceedings of the 30th ACM international conference on Design of communication Pages 105-114.
 - [11] A. Katasonov, " Lecture 6: Requirements Validation and Verification: Requirements Management and Systems Engineering " (ITKS451), Autumn 2008 University of Jyväskylä.
 - [12] L. He, Dr. Jeffrey, C. Carver, Dr. Rayford B. " Using Inspections to Teach Requirements Validation" , Vaughn Mississippi State University January 2008
 - [13] P. Heymans, " Some thoughts about the animation of formal specifications written in the ALBERT language" , Proceedings of the Doctoral Consortium of the 3rd IEEE International Symposium on Requirements Engineering (RE'97), Annapolis, MD, USA, January 6-10, 1997.
 - [14] G. Kotonya, I. Sommerville, " Requirements Engineering Processes and Techniques", John Wiley & Sons, England, 1998
 - [15] G. Kotonya , I. Sommerville, " Requirements Engineering". John Wiley and Sons, 2004.
 - [16] I Sommerville, " Integrated Requirements Engineering: A Tutorial", IEEE Computer Society, 2005.
 - [17] F. Fabbrini, , M. Fusani, , V. Gervasi, , S. Gnesi, , S. Ruggieri, " Achieving Quality in Natural Language Requirements " Proceedings of the 11th International Software Quality Week (1998).
 - [18] A. Stylianou, R. Kumar, M. Khouja, " A Total Quality Management-based Systems Development Process" The DATA BASE for Advances in Information Systems, Vol. 28, 1997, No. 3, pp. 59-71.
 - [19] H. In, D. Olson, T. Rodgers, " Multi-criteria Preference Analysis for Systematic Requirements Negotiation", In Proceedings of the 26th Annual International Computer Software and Applications Conference (COMPSAC'02), Oxford, England, 2002, pp. 887{892. IEEE Computer.
 - [20] P. Chuang, " Combining the Analytic Hierarchy Process and Quality Function Deployment for a Location Decision from a Requirement Perspective", The International Journal of Advanced Manufacturing Technology, Vol. 18, 2001, No. 11, pp. 842{849.
 - [21] V. Bouchereau, H. Rowlands, " Quality Function Deployment: The Unused Tool", Engineering Management Journal, Vol. 10, 2000, No. 1, pp. 45-52.
 - [22] D. Norman, "The design of everyday things. MIT press (1998).
 - [23] CREWS (ESPRIT N°21.903), Cooperative Requirements Engineering With Scenarios, <http://sunsite.Informatik.RWTH-Aachen.DE/CREWS>.

Interactive Image Search for Mobile Devices

Komal V. Aher
ME Student, Department of Computer,
Sinhgad Institute of Technology,
Lonavala, India.
aher.komal@gmail.com

Sanjay B. Waykar
Department of Computer Engineering,
Sinhgad Institute of Technology,
Lonavala, India.
sbwaykar@gmail.com

Abstract— now a day's every individual having mobile device with them. In both computer vision and information retrieval Image search is currently hot topic with many applications. The proposed intelligent image search system is fully utilizing multimodal and multi-touch functionalities of smart phones which allows search with Image, Voice, and Text on mobile phones. The system will be more useful for users who already have pictures in their minds but have no proper descriptions or names to address them.

The paper gives system with ability to form composite visual query to express user's intention more clearly which helps to give more precise or appropriate results to user. The proposed algorithm will considerably get better in different aspects. System also uses Context based Image retrieval scheme to give significant outcomes. So system is able to achieve gain in terms of search performance, accuracy and user satisfaction.

Keywords— *Mobile visual search, Multimodal search, Quantization, Histogram, Mobile device, color space etc.*

I. INTRODUCTION

Now a days, more and more people use phones or other mobile devices as their personal concierges surfing on the Internet. Besides this trend, searching is becoming pervasive and one of the most popular applications on mobile devices [6]. As desktop computers and mobile devices having different user interfaces especially for input types as well as search interest of users using mobile device also different from that on desktop screens. Existing search alternatives for mobile users include text-based search and local map search. Furthermore, photo-to-search is becoming persistent as the expansion of the computer vision and content-based image retrieval. It allows the user to capture pictures using the in-built camera on the phone and then begin search queries about stuff in visual proximity to the user.

Visual (image and video) search is still not that much popular on the mobile phone as judge against with text search, map search, and photo-to-search. The main

concern why these image search applications are not popular on mobile device is small screen of device and the existing search applications do not absolutely accommodate to the mobile and local oriented user intention. So in such situations where irrelevant images spoil the results and ruin the user experience, visual-aided tools can mostly improve the relevance of search results and the user experience [3]. Let's consider such a scenario in which the user has no idea of the name of target but can only depict its particular appearance, that only with a scene or general picture in the user's mind, such kinds of searches are not easy under present text-based search condition. But with the assist of visual aids, search for images not only based on text but also based on image content, these tasks can be more expressive.

So, in the proposed work we facilitate an intelligent and interactive visual search on mobile phones by taking full multi-modal and multi-touch functionalities of mobile device. If the users have an image in their hand, they can use it directly as a query and find matching images in dataset. Otherwise users can easily formulate a composite image as their search query by naturally interacting with the phone through voice and multi-touch. So the system allows users to express their implicit and explicit search intent well. We have designed a multimodal image search system to carry out different types of queries from mobile phones and expressing user's information needs in a better way. For improved search contextual information also added to the system. As a result, a powerful image search system with visual aids is designed. The proposed system gives compliment to existing systems and will give user friendly interface and satisfactory results to user conceptual query.

II. RELATED WORK

A. Visual Search

Visual search is a type of perceptual task requiring attention that usually includes an active scan of the visual environment for a particular object or the target among other objects or the distracters. There are some existing systems as:

1) Voice Query:

As Typing is a tedious task on the phone no matter whether a tiny keyboard or a touch screen is used. Even though voice queries are available on some devices, still many cases that semantic and visual intent cannot clearly expressed by these descriptions for search [3]. However, the users usually have to accept some ideal images amidst much more irrelevant results. In such cases where irrelevant images spoil the results and spoil the user experience, visual-aided tools can largely boost the relevance of search results and the user experience.

As the speech recognition became mature, phone applications using speech recognition rapidly grows. The most popular application is Apple Siri [11], which combines speech recognition, natural language understanding and knowledge based searching techniques, with an ability to make a speech conversation with the phone and get information and knowledge from it. The user can ask the phone for anything by only speech and get multimedia answers.

2) Text based Search:

Traditional text-based search engines like Google and Bing are available on mobile devices. However, extensive text queries are neither user-friendly on phone, nor machine-friendly for search engine. The fact is that the mobile users use only 2.6 terms on average for search [6], which can hardly express their search intent.

3) Photo to search:

It allows the user to capture photos using the in-built camera on the phone and then initiate search queries about objects in visual proximity to the user. This advance offers various applications such as identifying products, comparison shopping, finding information about buildings, movies, compact CDs, real estate, print media, artworks, etc.[3]

Photo-to-search applications are become important on mobile phones. These techniques enable users to search for what they see by taking a photo on the go. As Google Goggles, Point and Find [8], and Snaptell [9] are good

examples. These applications look for the precise partial duplicate images in their database and provide the users with related information of the query images. But, the search is only available for some vertical domains, such as CD covers, products, landmarks etc., where the partial duplicate images of the query image have been indexed in their database.

4) Sketch based Query:

Hand-drawn sketches were used in Sketch-based image search to search for satisfied images. But with these systems it is very hard to express user intent and is difficult for users without drawing experience to use such application. [10] [13]

In [7] the authors build a Sketch2Photo system that uses simple text-annotated line sketch to automatically synthesize realistic images. However, their work focuses on image composing instead of image retrieval.

B. Concept Based Image Retrieval

There is a traditional approach to any form of image search, as concept-based image indexing. Also known as description-based or text-based image retrieval, this type of search refers to keywords, tags, captions, subject headings or natural language text for the indexing and retrieval of text-based images. For years now, SEOs and digital marketers have been optimizing images so that search engines like Google could understand and properly index visual content [14].

C. Content Based Image Retrieval(CBIR)

With CBIR, search engines analyze the visual content of the image (pixels) rather than the metadata. In this the idea of "content" may refer to colors, shapes, textures, or any other information that can be derived from the image itself. CBIR is gaining popularity because of the inefficiencies and limitations inherent with metadata-based image retrieval. Optimizing for text-based retrieval can be time consuming and create unintended ambiguities. However, until recently, many image retrieval systems, such as Google-image search, were exclusively text based [14].

D. Reverse Image Search

Reverse image search is a CBIR query technique that involves providing the search engine with a sample image to base its query on. Reverse image search allows

users to discover content that is related to a specific sample image, popularity of an image, and discover manipulated versions and derivative works. Different implementations of CBIR make use of different types of user queries. Examples include Google Image search and Tin Eye [14] [16].

III. PROPOSED SYSTEM

A system is designed to tackle problems faced by user while dealing with mobile device to do image search. Here one mobile based application is designed for an image searching with user friendly and game like interface to give more relevant results to user. The application is useful for users who can't express their search intend clearly.

In a proposed work, standard architecture is followed as client-server. Mobile device with android OS is treated as client and laptop or computer as Server. Wherein Image database is maintained and handled on Server by administrator. At Client side, there are four ways are given to initiate image search which is a combination of all existing systems. Proposed system overview is shown below which will simply do image search task with user friendly manner:

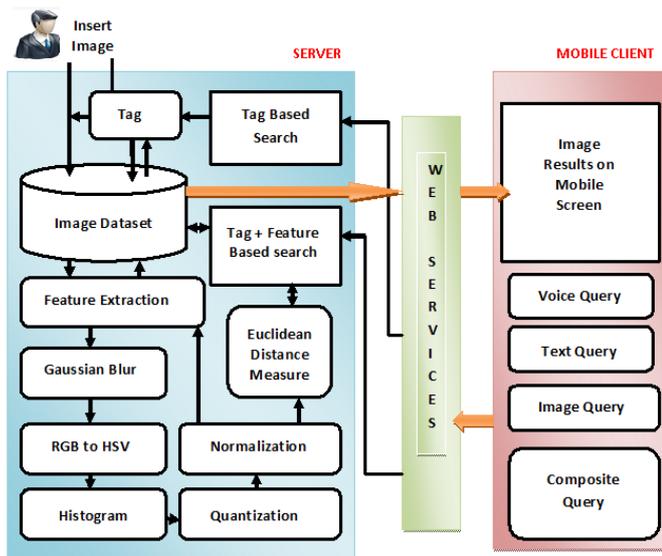


Fig. proposed System Architecture

A. Server Side Implementation:

The architecture basically describes on server side image dataset is created and maintained, wherein images are

stored as per categories by Admin. Two important things are considered while storing images into database- Image tag and its Features. So when any image is inserted into dataset at first tag is assigned to it and the features are extracted for each and every image.

Features of Images stored into the dataset are extracted. For that each image is passed through various image processing algorithms. Various operations are performed on each and every pixel of image such as Blurring, RGB to HSV, Color histogram calculation, Quantization and then Finally Normalization. All these algorithms are given as:

a. Image Blurring:

Blurring an image reduces the sharpening effects of image; which help to makes the detection more accurate. In blurring each pixel in the source image gets stretch over and mixed into surrounding pixels. Various steps for image blurring are:

1. Traverse through complete input image array.
2. Read individual pixel color value (24-bit).
3. Split the color value into individual R, G and B 8-bit values.
4. Calculate the RGB average of surrounding pixels and assign this average value to it.
5. Repeat the above step for each pixel.

Gaussian Blur is one of the most effective methods.

b. RGB to HSV Color Model:

After blurring pixels of image are converted from RGB to HSV model to make it easier to perform any operations. The HSV stands for the Hue, Saturation and Value, gives the perception illustration as per human visual aspect. Hue specifies the color type its range as 0 to 360. Saturation gives the vibrancy of the color ranges from 0 to 100%, and it also called as purity. Value, the brightness of the color ranges from 0 to 100%. RGB to HSV conversion steps are as follows:

1. Find minimum value of basic R,G,B

$$\text{Min}=\min(R, G, B)$$
2. Find maximum value of R, G, B

$$\text{Max}=\max(R, G, B)$$
3. Calculate HSV from RGB

$$\text{temp}=\text{Max}-\text{Min}$$

$$\text{Value}=\text{Max}$$
 If (Value==0) then

```

    Hue = Sat = 0
Else
    Sat=255 * (temp)/Value
    If (Sat==0) then
        Hue=0
    Else if (Max==R) then
        Hue= 0 + 43 * (G - B)/temp
    Else if (Max==G) then
        Hue= 85 + 43 * (R - B)/temp
    Else if (Max==B) then
        Hue= 171 + 43 * (R - G)/temp
If (H < 0)
H = H + 255
SetPixel.

```

It is observed that HSV model accuracy is higher as compared to RGB model. [12]

c. Histogram:

Color histogram represents the distribution of the composition of colors in the image. The histogram consists of bins where each bin defines a small range of pixel values. The value stored in each bin is the number of pixels in the image that are within the range. These ranges represent different level of intensity for each color component. The values in each bin are normalized by dividing with the total number of pixels in the image. Then, by counting the number of pixels in each of the bins, we get the color histogram of the image.

d. Quantization:

Quantization reduces the number of colors used in an image. While computing the pixels of diverse colors in an image, if the color space is oversized, then first segregate the color space into certain numbers of small intervals. Each of the intervals is called a bin. This process is called color quantization. The quantization of the number of colors into several bins is done in order to decrease the number of colors used in image retrieval.

e. Normalization:

Normalization is a process that changes the range of pixel intensity values. It sometimes called contrast stretching or histogram stretching. The purpose of normalization in the various applications is usually to bring the image, or other type of signal, into a range that is more familiar or normal to the senses.

f. Distance Measure Estimation:

Similarity measure is real value function that quantifies similarity between two objects, so here for similarity calculation between images Euclidean distance measure is used.

B. Client Side Implementation:

At another side on mobile device user can input his query in various manner. To communicate with the server, client should know the server IP. Basically there are four ways are provided to initiate image search:

a. Voice input:

To convert speech given by user as input is converted into text by using standard Google's speech recognition tool and result in more accurate text. As every android mobile have inbuilt speech to text conversion tool so proposed work uses the same.

Image result for Voice And Text Query:

- Input: Natural voice or text Query
- Output: Exemplar Images
- Algorithm:
 - 1: give text or voice query as input
 - 2: Tag based image search from database as query entity.
 - 3: arrange resultant images on mobile screen
 - 4: Output images as an exemplars.

As voice input is converted into text, that text is parsed and treated as tag. And according to tags image results are retrieved form predefined image dataset. And results as exemplars are shown to user.

b. Text input:

In a direct text input, user can give query as single keyword or multiple keywords. According to text input tag based images are retrieved and if resultant images are not specific to query then current results are forwarded as composite query of images and then tag and feature based image retrieval initiated then it results into more relevant images to the queries.

Image result for Voice And Text Query:

- Input: Text Query

- Output: Result Images
- Algorithm:
 - 1: Give text query as input
 - 2: Tag based image search from database as query entity.
 - 3: arrange resultant images on mobile screen
If (user is satisfied) then,
 Done
 Else
 Make composite query using multiple tags
 - 4: According to tags image collage is shown on mobile screen
 - 5: Image collage is given as query
 - 6: Feature based image retrieval at server
 - 7: Most similar results are shown on mobile screen
 - 8: Relevant image result.

c. Image input:

Image result for Voice And Text Query

- Input: single image
- Output: Similar images
- Algorithm:
 - 1: give image as input
 - 2: Feature Extraction by blurring, color conversion, histogram etc.
 - 3: Tag + feature based image retrieval started
 - 4: Similar results on screen.

d. Composite Images input:

In this user can select two image from results or pre stored image from device. Then that image is passed through image processing algorithms for feature extraction then feature based similarity is checked with help of Euclidean distance measure and finally relevant results are shown to user.

Composite Query Processing:

- Input: result images, photo
- Output: Final resultant images
- Algorithm:
 - 1: Check for input image or predefined photo or Composite visual query
 - 2: Segmentation Based Image representation using Histogram

- 3: Color feature extraction
- 4: Finding most similar images from dataset based on feature
- 5: Apply Euclidean distance measure
- 6: Rank or sort results
- 7: Final relevant output Images

IV. RESULTS AND DISCUSSION

For proposed image search application, results are calculated based on two parameters as time and accuracy. Time estimation for each type of search for proposed system is as follows:

Process	Time (ms)
Voice Search Activity	7168
Text Search Activity	1205
Image Input Result	5337
Image search Activity	1403
Composite Image search activity	1582

Table 1. Time Estimation

The accuracy is measured using precision and recall parameters. Some testing results are shown into below table:

Type	Actual Objects	Retrieved Objects	Correct Retrieved Objects
Image Search	11	8	5
Voice Search	11	2	2
Text+Image Search	11	7	6

Table 2: Searching Results

Using above values precision and recall is calculated as follows:

$$\begin{aligned} \text{Precision} &= (\text{Relevant Intersect Retrieved}) / \text{Retrieved} \\ &= \text{Correct Retrieved Object} / \text{Retrieved Objects} \\ \text{Recall} &= (\text{Relevant Intersect Retrieved}) / \text{Relevant} \\ &= \text{Correct Retrieved Object} / \text{Actual Objects} \end{aligned}$$

Type	Precision	Recall
Voice	0.625	0.454545455
Text + Image	1	0.181818182
Image	0.857142857	0.545454545
Total	0.827380952	0.393939394
Accuracy (%)	0.393939394	

Table 3: Precision and Recall Calculation

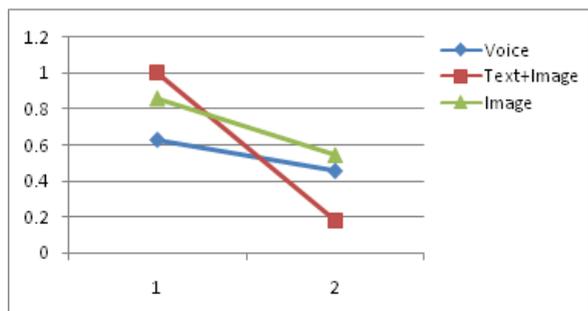


Fig. precision and recall for different inputs

The proposed system is working efficiently on mobile device and accommodate mobile screen with fully utilizing multitouch and multimodal functionalities of device. As per above results user will get satisfied with image results generated by proposed system.

V. CONCLUSION

The paper proposes a system which fully utilizing multimodal and multitouch functionalities of mobile devices and gives game like interface for image search. Thus, the paper gives system with ability to form composite visual query to express user intent more clearly which helps to give more specific or relevant results to user. The proposed algorithm will significantly improve in different aspects. System also use Context based Image retrieval schema to give relevant results. So system is able to achieve gain in terms of search performance, accuracy and satisfactory results to their imaginary queries.

ACKNOWLEDGMENT

I would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. I am thankful to the authorities of Savitribai Phule University of Pune. I'm also thankful to reviewer for their valuable suggestions.

REFERENCES

- [1] K. F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "An efficient and effective region-based image retrieval framework," *IEEE Trans. Image Process.*, vol. 13, no. 5, pp. 699–709, 2004.
- [2] B. Girod, V. Chandrasekhar, D. Chen, N. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai, and R.

- Vedantham, " Mobile visual search," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 61–76, 2011.
- [3] Houqiang Li, Yang Wang, Tao Mei, Jingdong Wang, and Shipeng Li, "Interactive Multimodal Visual Search on Mobile Device", *IEEE Transactions on Multimedia*, VOL. 15, NO. 3, April 2013.
- [4] C. Wang, Z. Li, and L. Zhang, " MindFinder: Image search by interactive sketching and tagging," in *Proc. Int. Conf. World Wide Web*, pp. 1309–1312, 2010.
- [5] Y. Wang, T. Mei, J. Wang, H. Li, and S. Li, "JIGSAW: Interactive mobile visual search with multimodal queries," *Proc. ACM Multimedia*, pp. 73–82, 2011.
- [6] K. Church, B. Smyth, P. Cotter, and K. Bradley, "Mobile information access: A study of emerging search behavior on the mobile internet," *ACM Trans. Web*, vol. 1, no. 1, May 2007.
- [7] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: Internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 124:1–124:10, Dec. 2009.
- [8] NOKIA Point and Find. Link: <http://pointandfind.nokia.com/>
- [9] Link: <http://www.snaptell.com/>
- [10] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang, " MindFinder: Interactive sketch-based image search on millions of images", in *Proc. ACM Multimedia*, pp. 1605–1608, 2010.
- [11] Link: <http://www.apple.com/iphone/features/siri.html>
- [12] Simardeep Kaur and Dr. Vijay Kumar Banga, "Content Based Image Retrieval: Survey and Comparison between RGB and HSV model," *International Journal of Engineering Trends and Technology (IJETT)* - Volume 4 Issue 4- April 2013.
- [13] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "An evaluation of descriptors for large-scale image retrieval from sketched feature lines," *Comput. Graphics*, vol. 34, no. 5, pp. 482–498, 2010.
- [14] Link : <http://www.google.com/>
- [15] Y. Cao, H. Wang, C. Wang, Z. Li, and L. Zhang, "MindFinder: interactive sketch-based image search on millions of images", in *Proc. ACM Multimedia*, 2010.
- [16] Link: <http://www.tineye.com/>
- [17] H. Xu, J. Wang, X. Hua, and S. Li, " Image search by concept map," in *Proc. ACM SIGIR*, 2010, pp. 275–282.
- [18] Zahariadis T., Daras P., Bouwen J., Niebert N., Griffin D., Alvarez F., Camarillo G., "Towards a Content-Centric Internet", *Towards the Future Internet - Emerging Trends from European Research*, IOS Press, pp. 227–236, Apr 2010.

Benefits of new laboratory tools in research and education

Gabriela Gladiola Andruseac, Mădălina Poștaru, Corina Cheptea, and Anca-Irina Galaction

Abstract— This paper aims to explore the possibility of using new laboratory tools like robots as an efficient training-educational tool for teaching technical subjects and not only these. It analyses the available solutions at the present moment which refer to the robots used in research and education. Characteristics, costs, strengths and weaknesses of some available solutions was compared and evaluated to introduce a guideline for an efficient method to support learning using educational robot. The advantage of “robotic revolution” is found in the potential of interaction with the system and creation of changes within its parameters. Development of educational robots is still in its early stages. To complete our understanding of educational robots, we should explore potential benefits of using robots for education, optimal design of them and any limits and challenges that must be addressed.

Keywords— educational robots, learning, education.

I. INTRODUCTION

THE economic development and technological inventions have led to the present situation of robots being the present society’s component, with an enormous potential in the educational system. Ten years ago, most of the robots operated in the field of industry, representing expensive investments with complex functions and capacities of finest execution and speed [1]. The present tendencies show that, in the near future, we will witness a real invasion of robots in the public places as well as private areas, the present statistics showing the fact that the number of robots is significantly increasing. If in 2002, at the worldwide level, the estimated number of robots was 4,5 million, from which 3,5 million operated in the services field, in 2010 the estimated number was at 8 million, from which 7 million would cover the same field [1].

Education as an essential activity in the evolution of a society could not have been neglected by the technological phenomena. The main idea of this paper is focused on the assumption that the use of educational robots could improve the school performance of pupils from elementary, secondary and high school. This assumption comes from the paradigm “*Tell me and I will forget. Show me and I will remember.*”

This work was supported by the Grant ERA-NET, ERA-IB “MICROTOOLS” authorized by The Romanian Executive Unit for Financing Higher Education, Research, Development and Innovation (UEFISCDI).

Gabriela Gladiola Andruseac, Mădălina Poștaru, Corina Cheptea, and Anca-Irina Galaction are within the Biomedical Science Department, Grigore T. Popa” University of Medicine and Pharmacy, Faculty of Medical Bioengineering, 9-13 M. Kogalniceanu Street, 700454 Iasi, Romania.

Involve me and I will understand” [2]. Moreover, teachers who have used this technology in the classroom state that there is a greater receptiveness of pupils regarding the information presented when they are supported by practical activities in which robots are involved. However, this assumption can also be exaggerated if it is not supported by a rigorous research and concrete results.

The main objective of our intervention was to identify the challenges from the educational robots field and to write a number of conclusions based on the most important and relevant research in the field. This way, the first part of this study is focused on the examination of specialized literature regarding the use of educational robots, while in the second part of the study; we have initiated a classification of the types of educational robots that are available at this moment on the market.

II. A REVIEW OF THE APPLICABILITY OF EDUCATIONAL ROBOTS

At the first stage of the development of this paper we have initiated a search related to the existence of a study connected to the situation of research about the use of robots as an efficient instrument in the educational process. As a result of this intervention there has not been identified any systematic study regarding this subject even though the variety of specialized literature offers numerous studies connected to the use of computers, games or mobile devices in the educational process. This way, the inclusion of the ICT field in various activities regarding the teaching, assimilation and assessment of pupils is largely debated with results which are already used regarding the positive impact of this field on the young generation. In this context, our study was made based by Kitchenham’s approach [3] with the aim of examining the present situation of research regarding the inclusion of educational robots in the curricular or extra-curricular area.

The objectives of the study have concentrated on:

- a) Identification of potential benefits following the use of robotics as an educational instrument in different areas of knowledge;
- b) The presentation of a synthesis of present studies regarding this approach of education;
- c) The definition of future research perspectives regarding the approach of the educational process from the point of view of supporting it with the help of robotics.

The examination of specialized literature was done by using

the international databases called IEEE Xplore, ScienceDirect and Springer Link, the search taking place by using the education word families: ‘to teach’, ‘to learn’, ‘to educate’, ‘school’, ‘robot’ and well known key words in the field of present robotics: NXT/EV3, Mindstorm, Arduino. The search has been restricted to papers published between 2004 and present. In table I we have presented the search protocol that was used in the interrogation of specified databases.

Table I. The protocol used in each database

Database / Field Information	Protocol
IEEE Xplore / Search on the field “Abstract”	(((((teaching or teach or learn or learning or education or educational)<in>ab) <and> ((robotic or robot or robotics or robots or lego)<in>ab)) <and> ((school)<in>ab)) <and> (pyr >¼ 2000 <and> pyr <¼ 2014))
Science Direct / Search on the fields “Abstract”, “Title” and “Keywords”.	pub-date >1999 and title-abstr-key((teaching OR learning OR teach OR learn OR education OR educational) AND (robotic OR robotics OR robot OR robots OR Lego) AND (school))
SpringerLink / Search on the field “Abstract”.	ab:((teaching or learning or education or educational) and (robots or robotic or robot or Lego or Mindstorm) and (school)) Content Type >Journal Articles Publication Date >Between Saturday, January 01, 2000 and Thursday, December 31, 2012

A. General key points of research regarding the use of robots as a tool in the support of the educational process

In order to respond to the stated objective of identifying the potential benefits connected to the approach of education which is supported by the robotics technology, we have concentrated on answering a question such as: ‘*To what extent, can the field of applicability of robotics be expanded?*’ *Is the technical science field preferred, while disregarding the non-technical ones? What curricular areas can be supported by the educational robots and which are the least approachable fields or less understood by pupils by using traditional methods? What competencies and abilities can be developed through this approach? Can we use the educational robots for ‘everybody’s’ education or just as an extra-curricular activity? Which is the role of robots in the teaching-learning-assessment triad and how does the role of teacher change in the context? We can also ask the question if the classical pedagogical theories find the applicability within this approach and which are the social and emotional applications of using educational robots?* All these questions refer to ‘what is being studied, when is studied and how is studied’. In this context we have identified and given examples of general points of the research in this field, the activity being supported by the specialized literature.

In order to respond to the questions launched above, there have been established the following criteria regarding the inclusion of papers which resulted after the interrogation of databases from our study:

1. The article presents the method of use of robotics as a teaching instrument and not the teaching of robotics;
2. The article presents the used educational robots in elementary, secondary and high schools;
3. The article presents a quantitative evaluation of learning based on the present recommendations in the field of education [4].

B. Which is the field of application of educational robots’ usage?

The first problem that was approached as part of the study was focused on the analysis of the applicability field of robotics technology, a field which can be divided into two big categories: technical education and non-technical education. Within this term of ‘technical education’ we find the robotics science, computer science and technology. After analyzing the results of the interrogation of databases, we observed that most of the studies (60%) had focused on the development of applications in the field of robotics: programming and construction of robots, automatic systems, mechatronics (Table II). With the help of sensors and actuators, robots are capable to explore and interact with the real world, enabling the appearance of new educational activities which could support the curricula topics in this field. In most of the cases these activities have as a goal the familiarization of students with technology and the introduction of programming notions [5], [6], [7]. In all these articles there are presented activities in which pupils build robots with their own hands. These activities have offered a strong feeling of ownership and interest for students.

The second field that has been observed in the area of educational robots belongs to mathematics [8], physics [9] or geometry [10]. In [9] an autonomous robot is described (Autonomous Educational Robot Mediator - AERM) which helps pupils understand physics, mathematics and kinematics. With the help of AERM pupils create abstract models of concepts and relevant properties from the real world. The study statistically demonstrate that pupils who have worked with AERM have significantly improved their capacity to understand relevant concepts from the course’s topics thus getting better results in tests than classmates who have not participated in robotics activities. In this field the movement of robots is the main characteristic that is explored for teaching lesson about Newton’s Laws, distances, angles, fractions, graphical constructions and interpretations [8], [10]. For example, in [8] students learn about rotations and transformations based on robots’ movement while in [11] and [12] science, technology, engineering and mathematics notions (STEM) are supported by the use of global positioning technology (GPS) and geographic information system (GIS).

Table II. Areas of application of educational robots (SpringerLink Databases)

Results	Discipline	No of Papers
45 Result(s) for "'educational robot" AND (lego OR mindstorm)' within 2000 - 2014	1. Computer Science	29
	2. Engineering	13
	3. Business & Management	1
	4. Earth Sciences & Geography	1
	5. Education & Language	1
316 Result(s) for "'educational robot" or "educational robotics" or "robotics in schools" AND (mindstorm, OR lego, OR robot)'	1. Computer Science	165
	2. Engineering	108
	3. Education & Language	17
	4. Physics	14
	5. Mathematics	12
28 Result(s) for "'robotics in schools'"	1. Computer Science	15
	2. Education & Language	6
	3. Engineering	4
	4. Business & Management	2
	5. Social Sciences	1

Other examples of non-technical applications of educational robots are found in [9] for kinematics area, in [13] for music orchestration or in [14] for teaching basic principles of evolution to secondary-school life science classes. An interesting field is found in [15], paper which significantly moves away from the field of technical and exact science by approaching the philological field. There are two goals of this paper: first of all, a definition is offered to 'science literacy' and this gives a framework to the analysis of relation between robotics activity and knowledge and competencies in the field of science literacy; and secondly, it reports the result of a study focused on how academically advanced students used science literacy skills to solve robotics problems and the learning gains they achieved as a result of participation in the robotics course.

The third field that has been approached by the specialized literature is represented by the field of teaching foreign languages. The implications of this approach are reflected in the students' attitude: they do not hesitate to speak a foreign language in front of a robot the way they do in front of a human trainer and during the repetitive actions of knowledge acquirement tiredness is not present on behalf of the teacher [16], [17], [18]. Typically, a working plan is:

1. the student is given a task: a picture is presented, he/she listens to a question, a story is narrated, watches a film;
2. the learner answers the question by writing, speaking or choosing an answer on the screen;
3. the system evaluates the answer and offers feedback;
4. the robot offers helpful advice and gets the students to reach the correct answer gradually.

An important aspect in this field is represented by the recognition of correct speech. That is why a human operator is used for such exercise, which controls the robot from 'behind the scene' [1].

The fourth field that has been identified is represented by assistive robotics where robots are used to develop cognitive abilities of children and teenagers [19].

In this field, an important place is occupied by the development of communication abilities of children with autism. In [20] it is done a comparative analysis of efficiency of LEGO therapy versus SLP (Social Use of Language Programme). Both methods are easy to implement for children between 6 and 11 years, with HFA (High Functioning Autism) and AS (Asperger Syndrome) but while the LEGO therapy uses a game based on collaboration, SLP uses methods of direct teaching. The same way, by using the LEGO teaching method and LEGO ROBOTICS, in [21] and [22] there are investigated from a statistical point of view, the effects of the implementation of these methods in the classroom by measuring the results got by pupils.

Moreover, studies show that the registered progress of autistic children or with learning difficulties are significant if robots are used in the teaching-learning-evaluation triad. As part of an experiment which took place in a school from Birmingham, the UK, there has been issued for the first time a new method of rehabilitation of autistic children; these children seem to respond better to stimuli and information which came from robots because these robots do not show emotions. There has been observed that pupils with autism – who have difficulties in social interactions and in expressing and understanding of emotions – perceived the robots as less threatening than their human teachers and as a consequence they interacted easier with them. Research in the field show that most of the time, autistic children are attracted by technology, get along well with computers and other technological systems and manage with their help to improve their communication and social skills.

C. Should we use robots in schools or outside schools?

The second point in the research regarding the use of robots as an efficient instrument to support the educational process is represented by the place where the training activities take place: in schools or outside schools. The activities done in school are curricular activities and are part of the formal

education. The activities done outside school are extra-curricular activities and are part of the non-formal education.

In the specialized literature we find examples of inclusion of robotics activities in the curricular area of formal education [21], [22], [23]. However, the effort of teachers to adapt to the new challenge given by the introduction of robotics in classes, has to be noticed [21], [22]. This thing indicates the necessity of an inter-disciplinary collaboration between the ‘technical’ and ‘non-technical’ teachers.

As it was expected, there are various examples from the extra-curricular area of robotics activities. These activities take place outside the classroom, as part of after-school programmes, in community organizations, museums, libraries or summer camps [24], [25], [26]. These activities are less formal but have the same educational result. It is important to have a friendly atmosphere for learning. The activities take place in locations in which pupils feel less intimidated or more comfortable than in a formal classroom. Analyzing the studies, we can observe that the cases in which robotics were applied as an extracurricular activity, always involved a ‘group of tutors’.

D. Should we use robots in schools or outside schools?

The second point in the research regarding the use of robots as an efficient instrument to support the educational process is represented by the place where the training activities take place: in schools or outside schools.

The activities done in school are curricular activities and are part of the formal education. The activities done outside school are extra-curricular activities and are part of the non-formal education.

In the specialized literature we find examples of inclusion of robotics activities in the curricular area of formal education [21], [22], [23]. However, the effort of teachers to adapt to the new challenge given by the introduction of robotics in classes, has to be noticed [21], [22]. This thing indicates the necessity of an inter-disciplinary collaboration between the ‘technical’ and ‘non-technical’ teachers.

As it was expected, there are various examples from the extra-curricular area of robotics activities. These activities take place outside the classroom, as part of after-school programmers, in community organizations, museums, libraries or summer camps [24], [25], [26]. These activities are less formal but have the same educational result. It is important to have a friendly atmosphere for learning. The activities take place in locations in which pupils feel less intimidated or more comfortable than in a formal classroom. Analyzing the studies, we can observe that the cases in which robotics were applied as an extracurricular activity, always involved a ‘group of tutors’.

E. At what age do pupils best benefit from robotics technology?

The use of robots for supporting primary education up to graduation at university has become, in the recent years, an interesting field for researchers in the education science all

over the world. Robots have become a popular educational tool in some middle and high schools in USA, Japan and some countries from Europe, as well as in numerous youth summer camps, raising interest in programming, artificial intelligence and robotics among students. Due to the youth’s preference shown towards technology, robots have been validated by researchers in the field of education as useful tools in teaching mathematics and physics [1].

Robotics is seen by many as offering major new benefits in education at all levels [27]. This impact of social robotics is even more crucial for children and teenagers, where robots can be used for their development and intellectual growth. Young people who are not interested in traditional approaches to robotics become motivated when robotics activities are introduced in classroom. Not all robotic kits will appeal to all kinds of students. For example, we cannot expect young children to build complex robots or even use them. On the contrary, to attract young children, the robot must have animated features like BeeBot robot [28]. The BeeBot is a colorful bug like robot that can move and it is suitable to teach mathematics and programming to young children.

In general, educational robots are designed to take into account the age and the requirements of the children (Table III). Different children are attracted to different types of robotics activities. In [29] it is presented a strategy for introducing students to robotics technologies and concepts, and argues for the importance of providing multiple pathways into robotics, to ensure that there are entry points to engage young people with diverse interests and learning styles.

Table III. Case studies across different background knowledge required of educational robots

Education Level	Educational Kits
Primary	Lego WeDo Robotics, Robotis Ollo, Thymio II, BeeBot
Secondary	Robotis Ollo, Lego Mindstorm NXT & EV3, Thymio II Robot, Fischertechnik Computing
Vocational Education	Fischertechnik Computing, Lego Mindstorm NXT & EV3, Thymio II robot
College, University	Robotis DARwIn-OP (Dynamic Anthropomorphic Robot with Intelligence), Lego Mindstorm NXT/EV3

F. Do pedagogical theories support the use of robotics in the instructive-educational process?

In this section, we discuss a few pedagogical theories that are the most prevalent within the domain of educational robotics. As the robotics technologies have developed, many researchers have tried to use robots to support and render efficient the educational process.

As seen from the Table IV, in the last 10 years there has been an explosive growth of interest in this area.

Table IV. Results for "educational robot" (2000 – 2014)

Databases	Period	No of papers
SpringerLink	2000 – 2005	11
	2005 – 2010	54
	2010 - present	60
IEEE Xplore	2000 – 2005	6
	2005 – 2010	16
	2010 - present	25
ScienceDirect	2000 – 2005	5
	2005 – 2010	11
	2010 - present	21

The use of robotics by non-engineering and non-technical instructors has been termed a “*robotic revolution*” [30]. The pioneers in this domain is connected to the name of Seymour Papert who proposed an approach to education called constructionism as opposed to instructionism – the traditional style of teaching [31]. In the constructionist approach, students learn from the design, assembly and manipulation of their own robots [32]. As a matter of fact, the foundation in the use of educational robots lies in the “*tell me, show me, let me do it*” approach [2].

The influence of robotics technology and the appearance of a new paradigm, like r-Learning, is studied at a global level by interdisciplinary teams of teachers, psychologists, sociologists, IT teachers and from the wide specialized literature we can conclude that the introduction of technology is associated with a number of changes in the area of educational process: the appearance of new roles, like ‘the learning of technology’, the rise of preoccupations in the political field regarding the responsibility for fields like the development of curricula and its adaptation to present demands; the formalization of curricula; opportunities for study that are more and more flexible when it comes to time and location; a change of principle related to the definition of ‘teacher’, who can be regarded as a facilitator, organizer of knowledge and producer of educational content; a feeling of insecurity among the educators about their lack of understanding and/ or competencies related to these new forms of teaching; a rising number of initiatives of academic development by an increased interest shown to new approaches in the methodology [1].

From a pedagogical point of view, the use of educational robots allows:

- d) the familiarization with the new technology that is omnipresent in any activity of the contemporary society;
- e) the acquirement of new knowledge and building competences in the fields like: electronics, mechatronics, IT, robotics and automation;
- f) the rise of motivation with the help of real time feedback and interactive solutions which stimulate the ability to understand and interpret, placing the student in the centre of his/her formation and keeping him/her permanently active;
- g) the improvement of logical rationalization by dividing a problem into smaller problems, logical organization of the judgment leading to a quick thinking and deeper knowledge of the problem;

- h) the optimization of learning conditions, the value of interactivity and interdisciplinarity;
- i) the support of independent learning;
- j) the development of practical training and integration of professional competencies;
- k) the development of a participatory strategy and encouraging personal development;
- l) the synergy between team work and individual work.

III. EDUCATIONAL ROBOTIC KIT – MINDSTORM EV3

The robotic platform based on Mindstorm EV3 kit contains hardware type elements: motors, gears, wheels, sensors and a microcontroller which is the ‘brain’ of the robot (Fig. 1).

**Fig. 1 Mindstorm EV3 Educational Robotic Kit**

The robot’s actions are generated by the software component that contains all the data-collection from sensors, analysis structure which gives the robot autonomous power to decide the types of actions that it has to do. These actions include: monitoring of parameters that came from sensors; storing collected data from sensors and transmitting it towards decision structures regarding the future actions of the robot; real-time Bluetooth transmission of data. Decision structures command the autonomous movement of the whole gear or just a few elements (an arm, a sensor, etc.) according to the collected parameters. Thus, the robot responds to sensor inputs and use outputs to control the environment.

Sensors are used for collecting data from various areas, data which is analyzed, processed and represent the support for making appropriate decisions. Basically, with the help of sensors, the robot ‘sees’ and ‘feels’ the environment in which it operates. The strong assets of the robotic platform are represented by the high level of control and environment analysis, artificial vision, flexibility and modularity of system.

IV. CONCLUSION

In the last 50 years, a change has happened in the work field when it comes to employment or removing manual jobs in favour of automatized services. Currently, the operation and control of production are based on automatized systems

(Automation Technology) in which mechanic, mecatronics, electronics and Information Technology play a very important role. Together they involve the development of new abilities and competencies that we need to acquire for the inclusion in the work market which is more and more competitive.

Cognitive research has confirmed that students learn more if they are involved in their study and interact with the educational material (study, experiment, and test). The advantage of the new robotics technology is found in the potential of interaction with the system and creation of changes within its parameters. Statistics reports show that robots could help pupils to acquire mathematical, language and team problem solving abilities.

This study has shown that educational robotics have an enormous potential as a learning tool, including supporting the teaching of subjects that are not closely related to the Robotics field. All the education areas can be covered by the use of robots in order to facilitate teaching, learning and evaluation of information.

Development of educational robots is still in its early stages. To complete our understanding of educational robots, we should explore potential benefits of using robots for education, optimal design of them and any limits and challenges that must be addressed.

To conclude, our message is that we do not intend that robots replace human teachers but highlight the added value that robots can bring to the classroom in the form of a stimulating, engaging and instructive teaching aid.

ACKNOWLEDGMENT

This work was supported by the Grant ERA-NET, ERA-IB “MICROTOOLS” authorized by The Romanian Executive Unit for Financing Higher Education, Research, Development and Innovation (UEFISCDI).

REFERENCES

- [1] Andruseac, G., Iacob, R., 2013, Exploring the potential of using educational robotics as an effective tool to support collaborative learning, 4-th IEEE International Conference on E-health and Bioengineering - EHB 2013, ISBN: 978-1-4799-2372-4, pp. 1-4.
- [2] Moradi, H., Bahri, A., 2004, The Use of “Tell me, show me and let me do it” in Teaching Robotics, Proc. AAAI 2004, pp. 160-164.
- [3] Kitchenham, B., 2004, Procedures for performing systematic reviews. Joint technical report Software Engineering Group, Keele University, United Kingdom and Empirical Software Engineering, National ICT Australia Ltd, Australia.
- [4] Kirkpatrick, D.L., Kirkpatrick, J.D., 2006, Evaluating training programs: The four levels (3rd ed.), Berrett-Koehler Publishers.
- [5] Balch, T., Summet, J., et al., Designing personal robots for education: hardware, software, and curriculum, IEEE, Pervasive Computing, 7(2), 2008, pp.5-9.
- [6] Mubin, O., Bartneck, C., et al., Improving speech recognition with the robot interaction.
- [7] Chiou A., 2004, Teaching technology using educational robotics, Queensland University, pp. 9-14.
- [8] Highfield K., Mulligan J., Hedberg J., 2008, Early mathematics learning through exploration with programable toys, Proc. Joint Conference Psychology and Mathematics, 2008, pp. 17-21.
- [9] Mitnik, R., Nussbaum, M., Soto, A., 2008, An autonomous educational mobile robot mediator. *Autonomous Robots*, 25(4), pp. 367-382.
- [10] Tertl, Studos. Tertl Studos, www.tertl.com.
- [11] Nugent, G., Barker, B., Grandgenett, N., 2008, The effect of 4-H robotics and geospatial technologies on science, technology, engineering, and mathematics learning and attitudes. In J. Luca, & E. Weippl (Eds.), *Proceedings of world conference on education*.
- [12] Nugent, G., Barker, B., Grandgenett, N., Adamchuk, V., 2009, The use of digital manipulatives in k-12: robotics, GPS/GIS and programming. In *Frontiers in education conference, 2009. FIE '09. 39th IEEE*, pp. 1-6, pp. 18-21.
- [13] Han J.H., Kim D.H., Kim J.W., 2009, Physical learning activities with a teaching assistant robot in elementary school music.
- [14] Whittier, L. E., Robinson, M., 2007, Teaching evolution to non-English proficient students by using lego robotics. *American Secondary Education*, 35(3), pp.19-28.
- [15] Sullivan, F. R., 2008, Robotics and science literacy: thinking skills, science process skills and systems understanding. *Journal of Research in Science Teaching*, 45(3), pp. 373-394.
- [16] Kanda, T., Hiran, o T., Eaton, D., Ishiguro, H., 2004, Interactive robots as social partners and peer tutors for children: a field trial, *Human-Computer Interaction*, 19(1), 2004, pp. 61-84.
- [17] Han J., Kim D., 2009, R-Learning services for elementary school students with a teaching assistant robot, *Proc. HRI, 2009*, pp. 255-256.
- [18] Chang, C.W, Lee, J.H., et al., 2010, Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school, *Educational Technology and Society*, 13(2), 2010, pp. 13-24.
- [19] Tapus, A., Mataric, M.J., Scassellati B., 2007, Socially assistive robotics, *IEEE Robotics and Automation Magazine*, 14(1).
- [20] Owens, G., Granader, Y., Humphrey, A., Baron-Cohen, S., 2008, *Journal of Autism and Developmental Disorders*, 38(10), pp. 1944-1957.
- [21] Hussain, S., Lindh, J., Shukur, G., 2006, The effect of LEGO training on pupils' school performance in mathematics, problem solving ability and attitude: Swedish data. *Journal of Educational Technology and Society*, 9(3), pp. 182-194.
- [22] Lindh, J., Holgersson, T., 2007, Does lego training stimulate pupils' ability to solve logical problems? *Computers & Education*, 49(4), pp. 1097-1111.
- [23] Balch, T., Summet, J., Blank, D., et al., 2008, Designing personal robots for education: hardware, software, and curriculum, *IEEE, Pervasive Computing*, 7(2), 2008, pp. 5-9.
- [24] Riedo, F., Retornaz, P., Bergeron, L., et al., 2012, A two years informal learning experience using the thymio robot.
- [25] Barker, B. S., Ansoerge, J., 2007, Robotics as means to increase achievement scores in an informal learning environment. *Journal of Research on Technology in Education*, 39(3), pp. 229-243.
- [26] Williams, D., Ma, Y., Prejean, L., Lai, G., Ford, M., 2007, Acquisition of physics content knowledge and scientific inquiry skills in a robotics summer camp. *Journal of Research on Technology in Education*, 40(2), pp. 201-216.
- [27] Johnson, J., 2003, Children, robotics and education, *Proceedings of 7th international symposium on artificial life and robotics*, Vol. 7, pp. 16-21, Oita, Japan.
- [28] Janka, P., 2008, Using a programmable toy at preschool age: why and how? *Proceedings SIMPAR, 2008*, pp. 112-121.
- [29] Rusk, N., Resnick, M., Berg, R., Pezalla-Granlund, M., 2008, New pathways into robotics: strategies for broadening participation, *Journal of Science Education and Technology*, 17(1), pp. 59-69.
- [30] Hendler, J., 2000, Robots for the Rest of us: Designing Systems out of the Box, In: Druin, A., Hendler, J. (eds.) *Robots for Kids: Exploring New Technologies for Learning*, pp. 2-7.
- [31] Papert, S., 1993, *The Children's Machine: Rethinking School in the Age of the Computer*, Basic Books, New York, 1993
- [32] Li L.Y., Chen G.D., 2009, *Researches on Using Robots in Education, Learning by Playing. Game-based Education System Design and Development Lecture Notes in Computer Science*, Volume 5670, 2009, pp. 479-482.

A Probabilistic Clustering-based Adaptive Histogram Thresholding Method for Fast Segmentation of Color Images

Abolfazl Mirkazemy
Department of Computer Engineering
Faculty of Engineering
Shahid Chamran University
Ahvaz, Iran

S. Enayatollah Alavi
Department of Computer Engineering
Faculty of Engineering
Shahid Chamran University
Ahvaz, Iran

Gholamreza Akbarizadeh
Department of Electrical Engineering
Faculty of Engineering
Shahid Chamran University
Ahvaz, Iran

Abstract— In this paper, a new color image segmentation method based on adaptive histogram thresholding and probabilistic clustering of random local samples is presented which is able to be used in parallel processing. Histogram thresholding methods are well known for their fast results and low time complexity, but nearly all thresholding methods work only over gray scale images and suffer from large deal of data loss during color space transformation to gray scale. In recent years, several approaches try to improve the accuracy of thresholding methods by adding the difference dimensions like entropy, probability function or pixel neighborhood to gray scale. Proposed method uses color channel components as main features of segmentation and then tries to find correlation between histogram's peaks and plains between each color components by using predefined random samples of image. This will be useful to define main color components of each object in image and then it will be easy to separate objects and background from each other with parallel processing. Low time complexity and ability of using parallel processing with high accuracy make this method a robust and reliable approach for color image processing in both RGB and HSI color spaces. The experimental results on benchmark datasets demonstrate that the proposed method is more efficient than the standard color image segmentation methods.

Keywords— adaptive histogram thresholding, color image segmentation, histogram plain, local sampling, parallel processing, peak detection, probabilistic clustering.

I. INTRODUCTION

Image segmentation is the first major step in any image processing algorithm or method, because all other preprocessing steps, like image normalization and denoising, are done in nearly same way. Image segmentation is a process that mainly separates an image into two major parts named background and foreground (our objects from each other). In recent years, color images have become very popular in many aspects of technologies due to their ability to provide more information than gray scale images because they are able to assign more details to image with meaningful colors like Magnetic Resonance Imaging (MRI), Synthetic Aperture Radar (SAR), and many other aspect of since. Hence, many old image segmentation methods are developed for gray scale images and are unable to detect of detach colors from each other's. The only way to use these methods is color space transform. This transformation is very costly process due to the large amount of data loss in both color transforms from three or more dimensions of color space to gray scale and from gray scale to main color space. Many researchers have developed

color image segmentation methods by using color features as discrimination measures. But, these methods have a serious problem which makes them very hard to be used in real case and they have also large time complexity.

Two major subjects most be defined to develop a new method for color image segmentation. First, a suitable color space should be selected. Choosing an effective color space is useful to present an image with its entire details. It also provides good conditions for extracting good features that have enough discrimination power in segmentation process. Second, a powerful similarity or dissimilarity measure from extracted features should be found. This measure and the way to use that, provide a large variety of segmentation methods.

The paper is organized as follows. In section II, color space selection is introduced and major segmentation methods have briefly reviewed. In this review, the latest approach in last decade with special focus on thresholding methods have investigated. Then, the proposed method based on adaptive histogram thresholding is shown in section III. The experimental result is presented in section IV. Section V draws some conclusion. Finally, the future works are shown in section VI.

II. Related works

There are two common views in image processing and machine vision about color and color spaces. First, one lays on idea that each color can be obtained by mixture of primary colors (Red, Green and blue) or secondary colors (Cyan, Magenta and Yellow). On the other hand, second view defines color as metric parameters of energy and frequency. In this view, any color will obtained by triplex of Radiance, Luminance and Brightness. So, by these views, two main types of color spaces can be modified: intensity-base colored spaces, like RGB and CMY, and energy-base colored spaces, like HSI and HSV. In this paper, both RGB and HSI colored spaces have used as candidates of both colored spaces. Using color-base colored spaces are more easily and faster due to the color component which are less depended to each other and there is straight relation between pixel value and color components. Energy-base colored spaces are more complex because they are depended on each other. For example, colors in the HSI space have obtained by following formulas:

$$\text{chromaticity} = \text{hue} + \text{saturation} \quad (1)$$

$$\text{color} = \text{chromaticity} + \text{intensity} \quad (2)$$

Each color space have his own advantages and disadvantages that have difference usages for example RGB color space mainly used in digital cameras and monitors and

HSI color space used in color evaluation and digital processing. Other color spaces categories of color space can be found in [1], [2], and [3].

There are many classification tasks for image segmentation methods. According to [4], these methods can be categorized into six main categories as follows:

1. Region based segmentation:

In these methods, we split image into several sub images and then use three primary rules (region merging, region splitting and region growing) over them in iterative steps, after several iterations images well converged into some regions that each one will contain an object in main image. The most advantage of these methods are their tolerance to noise but with cost of large time complexity. These types are explained by ref. [5] very well.

2. ANN¹ based segmentation:

ANNs are popular algorithms in both computer and electronic science because they are fast, reliable and easy to use. In this type of segmentation each pixel will be assigned to a neuron in first ANN layer and ANN will trained with pre segmented images. Intra neuron weight will be modified by pixel neighborhood correlations. The most used ANNs for image segmentation are SOM, Hopfield, FFNN and PCNN. [6] And [7] are good example of this category

3. Fuzzy theory based segmentation:

Fuzzy logic and Fuzzy sets are new developed part of knowledge that their usage are increasing each day. In these type of methods we change our view about pixel feature as bimodal color and gain closer view to real world colors by transform image through fuzzification function and provide more information about pixel features and its neighborhood. These methods are more accurate than Crisp based methods and have better results especially in noisy images. We are also able to use FCM and Fuzzy K-means for clustering [8]. Many of Fuzzy segmentation methods will use mixed with morphological methods like [9].

4. PDE² based segmentation:

PDE models and PDE equations are use active contours or snake transforms to segment image objects by transforming segmentation problem into PDE. Snake transform put small geometrical shapes in random parts of image and iteratively increase its sides as its covered area grows larger and finally converged into edges of objects. The high performance of this category makes them very popular in image segmentation. The most famous algorithms of this category are Mumford shah, Level Set and snake transform [10]. [11] Will present interesting method for medical image segmentation.

5. Edge based segmentation:

These types of methods focus on intensity changes in pixel neighborhood, by moving a neighborhood window (usually 8 or more) around each pixel and looking for meaningful intensity changes and detecting edge of objects in image [12]. There are several measure for detecting changes in neighborhood like gray histogram threshold, Gradient, zero crossing [13] and Laplacian of Gaussian (LoG) [14].

6. The last category is thresholding based segmentation:

These types of methods are the oldest methods and most used methods for image processing. The main idea of these methods is to find a suitable place in image histogram that can extract objects from background. This category contain fastest methods for image segmentation.

There are several ways to categorize thresholding methods, here we talk about two popular categories. First one is hard thresholding in front of soft thresholding and second view is local thresholding again global thresholding.

1. Hard thresholding: in this method manually or by predefined rules an exact point over image histogram will be chosen and sets as threshold. By assumption of each histogram present an object in image, point (T) will calculated in area between peaks borders and if peaks have overlap the deepest point in valley will be chosen.
2. Soft thresholding: this type of threshold will be chosen adaptively based on unique information of image during several iterations. Many famous methods like ANN, Generic or other evolutionary algorithm, FCM, K-means and Mean-shift will used in iterations to adapt threshold into an exact point.

According to [15], thresholding methods can be classified into two classes as below:

- a. Local thresholding: in this approach image will divided into several sub images and one threshold for each of them will calculated. After that thresholds will combined by some methods like averaging, median or other ways and one final threshold calculated for entire image.
- b. Global thresholding methods can be explained in two major classes.
 - I. Region depended Global thresholding: these type of methods use pixel neighborhood and a probability function around central pixel to define threshold for images methods like Histogram transform, Gradient relaxation and Dreavi & pal method are from this category.

¹ Artificial neural networks

² Partial Differential Equation

II. Point depended Global thresholding: simple idea of clustering or classifying image's pixel into several class is core of these methods and very famous methods like OTSU, entropy, minimum error and concavity analysis will be classified in this category. Main difference between these methods are measure of clustering /classification like inter class variance, intra class variance, Kapur entropy [16] or anything else.

III. Proposed method

The main idea of this method developed over ability to process each color channels of color space independently and paralleled. In fact in this paper we try to take advantage of divide and conquer method to solve the image segmentation problem and reduce the dimensions and complexity of problem. Hence it is possible to lock at a color channel separately as single dimension of main image that contain depended, semi depended or independent information of image pixels, we can use histogram thresholding method over it without using of information in other channels. In this situation we are able to done same process over remaining color channels easily and in the end it just needs a good method to combine the independent results of each color channel into one segmentation rules for whole color space and entire image.

Another innovation which is used in this paper is conditional accepting plain part of histograms as independent objects. One of must problems that makes thresholding methods inaccuracy are inability to segment objects with continues color spectrums (like a hat in sun light). These object are shown in histogram as flat continues parts (these object did not have enough pixel density to create a peak). Unfortunately nearly all of histogram thresholding methods are working only in one part of histogram peak or valley, but histogram his third part named plain. Plains are usually didn't have useful or effective information for segmentation but in special condition they are able to be counted as one independent object. These conditions are minimum length and under the curve area. We will explain these conditions more in conclusion part beside other confidence measures.

This method include 10 steps and only first and final steps we are need to use main color image color space. Figure 1 shows the proposed methods flowchart.

In first step, input image will be read and its color channels separate its' from each other (here RGB color space is chosen but we are able to done same process over other color spaces too), then in second step sampling algorithm starts divide image into several sub images and took random samples from each one.

One of must important parameters of this method is sampling rate because it directly effect on time complexity and accuracy of method. Sample rate define how many sub images must be created and how many samples must be taken in next steps many of processes act directly or indirectly with this sampling rate.

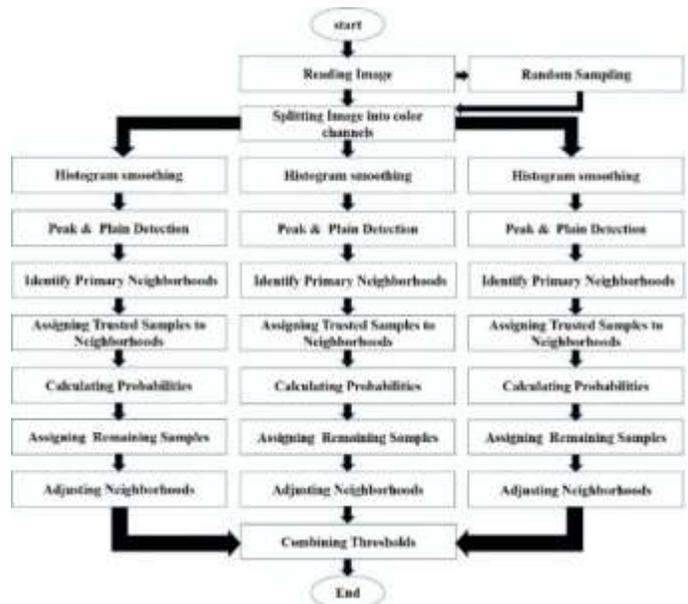


Fig. 1. Proposed method flow chart

If sampling rates chosen too high time complexity will be increased exponentially and in the other hand if it chosen so low the entire methods error increased and it even might led to wrong and unconvargable results.

While sampling algorithm start sampling third step start Parallelized. In other treads of program each color channel of image's histogram extracted and histogram smoothing method are applied on them. The main reasons of histogram smoothing are:

1. Normalizing histogram for peak detection step.
2. Increasing peak detection accuracy by reducing the probability of detecting small peaks or too close peaks³
3. In plain detection it is very important to have smooth and flat parts of histogram. These plain are usually have lots of inequality and small peaks and valleys that makes plain detection hard. Smoothing algorithm make this step very easier and faster.

Next step is plain and peak detection. In this step first we use differentiate of histogram curve to find extreme points, then use sign algorithm to classify extreme points into peak and valley classes. In this method we use peaks instead valleys (valleys are more common measure than peaks in histogram thresholding methods). The aim of this choice is gain the ability of moving toward the histogram peaks to detect plains (valleys shows the exact point of threshold and objects borders but peaks shows the objects pixels). We use some confidence measure to ensure that right peaks are chosen for segmentation. These confidence measure are minimum distance between two peaks and minimum acceptable peaks height. After that an algorithm start search for finding plains. In this algorithm first of all minimum swing rate of plain area

³ Histogram smoothing try to merge peaks that are too close to each other. If two or more peaks of histogram are too close to each other(less that 10 levels for example) they are defining one object in image and must be seen as one bigger peak.

must be defined, this measure used in histogram pacing because even with histogram smoothing plain parts still have lot of inequality, this measure help us find plain with acceptable swings. We also use confidence measures of minimum length and under curve area.

In fifth step, primary neighborhood around each peak will be defined (it will be called trusted or secure zone) and use plains length as their own neighborhoods. Then in next step samples whom gathered by sampling algorithm will be used. And find how many of these samples are in neighborhood of each interested area.

In seventh step, we try to define prior probability for each interested area based on two measure:

First measure is number of samples that are in primary neighborhood of area. As we choose small neighborhood around each area we can ensure that this sample belong to this area of interest.

$$\text{prior probability} = \frac{\text{number of samples in neighborhood}}{\text{number of all samples}} \quad (3)$$

Second measure is under curve area or peak's pixel density. This measure explain importance of number of pixels under of each curve. in fact this measure is auxiliary evidence of first measure. As height as height a peak is or as long as a plain is, shows how much pixels might belong to that area in main image. This fact is counted in this measure to support samples and make segmentation more accure.

$$\text{pixel density} = \text{neighborhood distance} \times \text{curve height} \quad (4)$$

Eighth step we assign remaining samples to interested areas by using maximum probability from previous measure and Euclidean distance between sample each interested area centers. Sample belong to area that have maximum P measure. In fact this step perform clustering algorithm with statistic cluster heads and probability assignment. In our experience we use some other methods. Some of them suffer from poor accuracy and some of them was too slow to be used in fast methods we will name some of them in conclusion section.

$$P_{\text{total}} = \text{probability measures} / \text{Euclidean distance} \quad (5)$$

Last parallel step is adjusting primary neighborhood with counting min and maximum samples amount that assigned to each neighborhood. In this step we calculate the minimum and maximum color level that assigned to each cluster, and define them as new threshold around each interested area.

In The final step we combine the results of each color channel and combine them into one result that could be used for segmentation of entire image at once. In this step we can assign each pixel to one cluster without using its neighborhood pixels information.

IV. Results

We use presented method over famous image (house) that used for color image segmentation many times and become one of most popular and well known grand truth for comparison and evaluation of segmentation effectively. This method was implemented in several condition and here we present one of our best results. In Implementation phase parameters that shown in table 1 are set for images.

Table 1. Experimental result's parameters

parameter	amount
Plain swing rate	4%
Minimum peaks height	5% ⁴
Minimum plain height	0.8%
Primary peaks neighborhood	20 level
Minimum plain length	30
Smoothing rate	0.1
Sampling rate	400
Minimum peaks distance	35

These parameters are extracted from post experiences and defined after using those result in curve to estimate the right value by estimation algorithms.

Figure 2 shows the input image and its sub images with random samples that shown with black color.

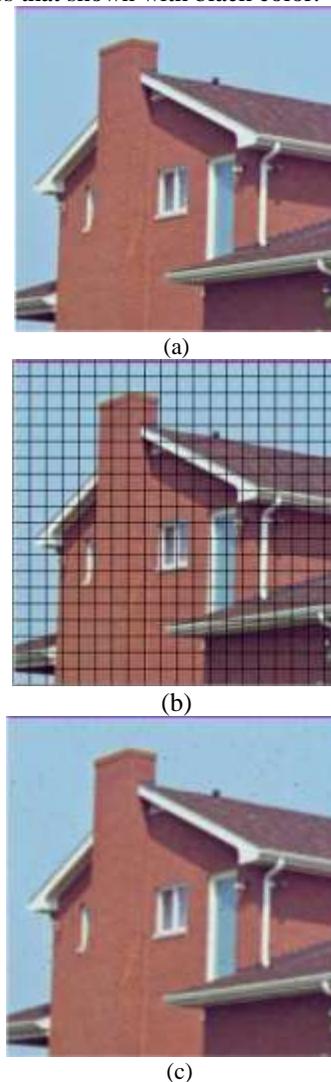


Fig. 2. (a) Primary images, (b) Sub images for sampling, (c) Random samples. Black points present random samples location

⁴ At least five percent of entire pixels of image must be under exact location of peak

After that each color channel processed separately. The results of blue channel histogram smoothing is shown in figure 3

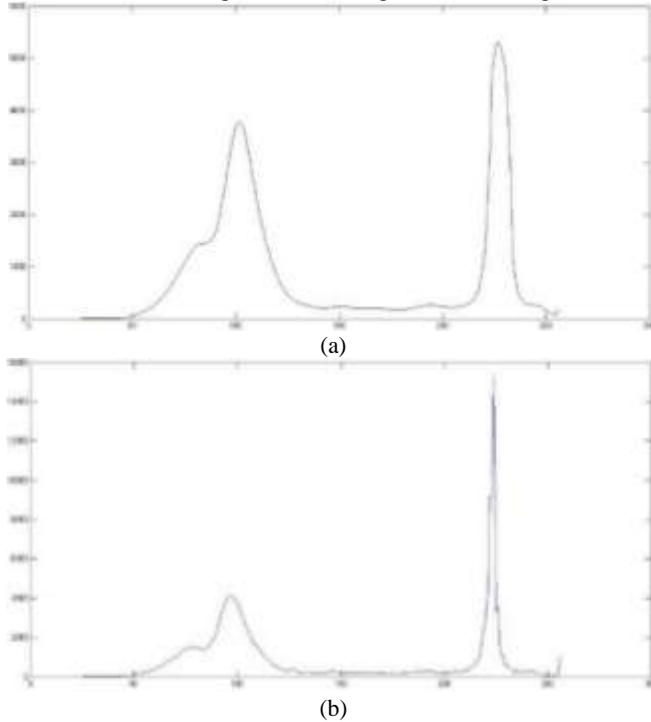


Fig. 3. Image's blue component histogram. (a) Original histogram (b) Histogram after smoothing

As you can see there are three areas of interest might be pointed in image. Two peaks and one plain areas is identified and used in next processes. Figure 4 shows interested area of Blue channel.

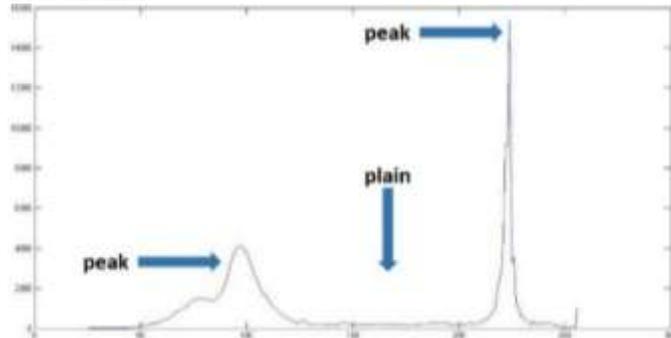


Fig. 4. Histograms interested areas

From 400 random samples that taken in sampling step 164 samples belong to first peak, 29 samples belong to plain and 132 samples belong to second peak. These numbers are divided to number of samples and used as prior probabilities. After this step a neighborhood defined around each area⁵. This area is use as secured zone (trusted zone) for first step of sample assignment. Peaks positions on histogram are 105 and 229. The secured zone of them are within range of [95-115] and [219-239]. Algorithm also detect a plain that meets the

⁵ Predefined parameters for each peaks neighborhood are 20 levels of gray scale (10 level each side) and for the plain are as long as detected plain.

acceptance condition in range of [135-205]. Figure 5 shows the interested area's neighborhoods.

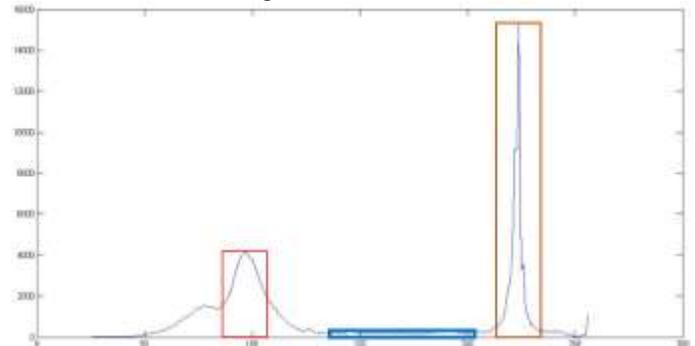


Fig. 5. Interested areas with predefined neighborhood (secured zones)

For calculating the curve density inside the each secured zone. The height of each zone center are multiplied to secure zone length. The result of this stage is 21.151.091, 141700 and 19038986. With these area space and probability we are able to step into next step and assign remaining samples to interest areas. These numbers and prior probabilities are multiplied with each other and divided to Euclidean distance of each sample. This parameter calculated for each interested area and in the end sample belong to the area that have maximum amount.

Then minimum and maximum samples that assigned to each are is calculated and these numbers are uses to adjusting the neighborhood of areas in this step primary neighborhood adjusted through table 2.

Table 2. Neighborhoods around each interested area before and after adjustment

	Primary neighborhood	Adjusted neighborhood
First peak	95-115	0-125
Plain	135-205	126-212
Second peak	219-239	213-255

In the end these result used with two other color channel (Red and Green channel) are combined. The neighborhoods are used as thresholds to divide color space into several subspace to segment image. In this experience two types of predefined color table are used to assign colors as lock up table to segmented areas. Figure 6 shows the results of method with two different color assignment to segmented areas.

V. Conclusion

In this section first we talk a two subjects that we pointed in proposed method section and then talk about conclusion. If an object in image want to be counted as independent object it must have least enough pixel density to considered as object. In image histogram must of pixel densities are concentrated around peaks and they are present the biggest objects in image (usually the biggest peak in image histogram present background). With this fact if we want to define a plain as object it must meet minimum density. Certainly it must be long enough and height enough in histogram to meet minimum expectations.



Fig. 6. Experimental results

Before these result will achieved some other methods were developed and implemented. That each of them have their own advantage and disadvantage. The first experience was using K-mean clustering over samples. This method's main problems were random sample inserting and dynamic cluster head. If first random samples was not close enough to the main peaks of histogram, final result converged to wrong points. In the other hand if we try to add selected sample first there is possibility to by adding out lair samplers dynamic cluster head incline to wrong area and after that accept samples that not belong to that cluster.

Another experience was testing two dimensions Euclidean distance between peaks top to each sample. This methods problem was inability to use over plain area of histogram. This method is known as one new and robust histogram thresholding named Triangle Method that explained in [17]. Unfortunately this method didn't perform well over plain area of image due the low height of plain's center area. Because of that these points nearly all times win the competition in comparison with high altitude peaks, and final threshold will converged to wrong point.

This methods design for fast (near real time) color image segmentation and aim for solid and static run time in first step of development. All of segmentation methods have same starting and ending phase reading image and applying

segmentation over areas of image. If we ignore these steps (that have same complexity for all methods), this methods time complexity (if we assume image size as compression parameter) is:

$$O = (\log_{\text{sample rate}} \text{image size})^2 \quad (6)$$

Because only in sampling phase image size effect on time complexity (in this step two counting rings provide sub images by dividing image size to sample rate). So it seems reasonable to conclude that this method's time complexity is nearly independent from its input image. But its complexity is depended to other parameters like histogram range (if we use RGB color space each histogram have 256 level and if HSV or HSI color space will used it have range between zero to one) that is fixed amount and will not count is time complexity calculation. Another parameter is sample rate that explained before and final effective parameter for time complexity is number of interested areas the detected in histogram, this measure is difference from image to image and there is no way to be predict such a parameter. But this parameter is not bigger than 4 or 5 and can easily be ignored.

In this paper we present new segmentation method for color images that is able to run paralleled, fast enough to be used in real time cases and don't convert image to gray scale (without data loss). It have acceptable accuracy and reliability, but it also have disadvantages like difficult parameter determination. Accuracy of this method is completely depended to its predefined parameters and must be determined carefully.

VI. Future works

Presented method is still needs a lots of work and optimization for getting ready and useable in devices and programs and plenty of tests must be passed to method's effectiveness and reliability will proved. One of optimizations that needs to be done is finding a relation between image and control parameters that makes method fully adaptive. In the other hand we also keep testing well known clustering algorithms and adapting them to our method for finding better results and accuracy. In same time we try to adapt this method for other color spaces like YIQ and NTSC.

This method right now is deployed over a mobile robot to aid in path planning of mobile robot and its path planning algorithm is under developing.

VII. References

- [1] Busin, Laurent, et al. "Color space selection for unsupervised color image segmentation by histogram multi thresholding." *ICIP*. Vol. 4. 2004.
- [2] N. Vandenhroucke, L. Macaire, and J.-G. Postaire "Color image segmentation by pixel classification in an adapted hybrid color space. Application to soccer image analysis," *Computer Vision and Image Understanding*, vol. 90, no. 2, pp. 190-216, 2003.
- [3] Y. I. Ohta, T. Kanade, and T. Sakai, "Color information for region segmentation," *Computer Graphics and Image Processing*, vol. 13, pp. 222-241, 1980.
- [4] Pal, Nikhil R., and Sankar K. Pal. "A review on image segmentation techniques." *Pattern recognition* 26.9 (1993): 1277-1294.
- [5] H. G. Kaganami and Z. Beij, "Region based detection versus edge detection," *IEEE Transactions on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 1217-1221, 2009.
- [6] B. J. Zwaag, K. Slump, and L. Spaanenburg, "Analysis of neural networks for edge detection," 2002.

- [7] C. Amza, "A review on neural network-based image segmentation techniques," *De Montfort University, Mechanical and Manufacturing Engg., The Gateway Leicester, LE1 9BH, United Kingdom*, pp. 1-23, 2012.
- [8] S. Naz, H. Majeed, and H. Irshad, "Image segmentation using fuzzy clustering: A survey," in *Proc. 6th International Conference on Emerging Technologies*, 2010, pp. 181-186.
- [9] I. Irum, M. Raza, and M. Sharif, "Morphological techniques for medical images: A review," *Research Journal of Applied Sciences*.
- [10] Sahoo, Prasanna K., S. A. K. C. Soltani, and Andrew KC Wong. "A survey of thresholding techniques." *Computer vision, graphics, and image processing*41.2 (1988): 233-260.
- [11] S. Bueno, A. M. Albala, and P. Cosfas, "Fuzziness and PDE based models for the segmentation of medical image," in *Proc. Nuclear Science Symposium Conference Record, IEEE*, 2004, pp. 3777-3780.
- [12] S. Lakshmi and D. V. Sankaranarayanan, "A study of edge detection techniques for segmentation computing approaches," *IJCA Special Issue on "Computer Aided Soft Computing Techniques for Imaging and Biomedical Applications" CASCT*, 2010.
- [13] B. Sumengen and B. Manjunath, "Multi-scale edge detection and image segmentation," in *Proc. European Signal Processing Conference*, 2005.
- [14] M Sharif, S Mohsin, M. Y. Javed, and M. A. Ali , "Single imageface recognition using laplacian of gaussian and discrete cosine transforms," *Int. Arab J. Inf. Technol.*, vol. 9, no. 6, pp. 562-570, 2012.
- [15] Sahoo, Prasanna K., S. A. K. C. Soltani, and Andrew KC Wong. "A survey of thresholding techniques." *Computer vision, graphics, and image processing*41.2 (1988): 233-260.
- [16] Kiani, Hamed, Reza Safabakhsh, and Ehsan Khadangi. "Fast recursive segmentation algorithm based on Kapur's entropy." *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on. IEEE*, 2009.
- [17] Zack, G. W., W. E. Rogers, and S. A. Latt. "Automatic measurement of sister chromatid exchange frequency." *Journal of Histochemistry & Cytochemistry*25.7 (1977): 741-753.

Big data analytics in prevention, preparedness, response and recovery in crisis and disaster management

Dontas Emmanouil¹, Doukas Nikolaos²

¹Hellenic Army, Artillery School, Nea Peramos, Greece

²Hellenic Army Academy, Vari, Greece

Abstract— The scientific area of crisis management has been in the center of attention for multiple disciplines especially the computer science. In the information centered and computer driven world, a major aim of computer scientists is to manage and analyze Big Data, extract information from heterogeneous sources and store it in unified structure formats that allow further processing. In this paper, Big Data analytics techniques and tools that are useful in all phases of crisis management are presented. Furthermore, a system-engineering approach of a big data management system will be analyzed that comprises of four phases; data generation, data acquisition, data storage, and data analytics. Benefits of the usage of Big Data for crisis management are analyzed. An innovative view of open problems concerning Big Data in crisis management is introduced.

Keywords—Big data analytics, crisis and disaster management.

I. INTRODUCTION

Effective management of crises and disasters, is a global challenge. All communities are vulnerable to crisis, both natural and induced by human activities. A systematic process with principal goal to minimize the negative impact or consequences of crises and disasters, thus protecting societal infrastructure, is called effective crisis and disaster management. It is imperative throughout the world to increase knowledge of crisis and disaster management, for the purpose improving responsiveness. All the above aims may be facilitated by Big Data Analysis.

Big Data and Computer Science

The concept of Big Data project is fundamentally related to computer science since the beginning of computing. The term Big Data describes amounts of data obtained with technological means that are normally unusable by humans due to volume and which with appropriate automated processing will extract actionable information. [1]

Big Data Characteristics

Big Data may be characterized as having four dimensions: Data Volume, measuring the amount of data available, with typical data sets occupying many terabytes. Data velocity is a measure of the rate of data creation, streaming and aggregation. Data variety is a measure of the richness of data representation – text, images, videos etc. Data value,

measures the usefulness of data in making decisions. [2]. A further characteristic has recently appeared, namely Variability, which represents the number of changes in the structure of the data their interpretation. Gartner [3] summarizes this in the definition of Big Data as high volume, velocity and variety information assets that demand cost effective processing.

Big Data in Crisis Management- Surveillance

The management of large volumes of data is perhaps one of the biggest challenges to be addressed by computer science. The wide variety of data acquisition sources available in times of crisis creates a need for data integration, aggregation and visualization. Such techniques assist crisis management officials to optimize the decision-making procedure. During the outburst of a crisis, the authorities responsible must quickly make decisions. The quality of these decisions depends on the quality of the information available. A key factor in crisis response is situational awareness. An appropriate, accurate assessment of the situation can empower decision-makers during a crisis to make convenient decisions, take suitable actions for the most affective crisis management [4]. Situational awareness definitions: “*perception*, where elements of the current situation are observed, *comprehension* where information obtained through observation is combined and interpreted and, *projection* where sufficient information and understanding exists to make predictions about impending events” [5]

II. A BIG DATA CHAIN

A. Big Data systems-engineering approach

A systems-engineering approach of a big data management system operates in four phases: data generation, data acquisition, data storage, and data analytics [6]. A big-data system is complex, providing functions to deal with different phases in the digital data life cycle, ranging from its birth to its destruction. At the same time, the system usually involves multiple distinct phases for different applications. [7]. Raw data can be taken as the raw materials with data generation and data acquisition being the corresponding exploitation process. In the same sense, data storage may be considered as a buffering process and data analysis as the final production process that utilizes the raw material to create new value [8]. The first stage leading to analysis is Data generation. The rate of data generation is increasing due to technological advancements. Indeed, IBM

estimated that 90% of the data in the world today has been created in the past two years [9]. The cause of the data explosion has been much debated. A related example is the huge amount of internet data being generated, such as internet forum posts, social media, chatting records. This huge amount of data may be unusable, but via suitable analyses may yield useful information concerning the habits and hobbies of users. Analyzing this information may render possible to predict behaviors, feelings and trends.

The data generation process is also subject to study, as it comprises of both controlled and unpredictable components. A set of sensors deployed in order to observe a particular situation, is a controllable source of high volume data. On the other hand, there exist in the internet large numbers of users, each one bestirring themselves independently and generating independent data traces. These data traces, when viewed as a total may provide information with serious implications for the economy, the defense and other topics of interest. Hence, the term big data is designated to mean large, diverse, and complex datasets that are generated from diverse data sources, both physically and virtually distributed, that include sensors, video, click streams and many other sources. [10].

B. Data acquisition

Data acquisition refers to the process of obtaining information and is subdivided into data collection, data transmission, and data pre-processing. One of the aims of the data acquisition phase is to aggregate information in a digital form for further storage and analysis. Firstly because data may come from a diverse set of sources, such as websites that host formatted text, images and videos, etc. Data collection refers to dedicated technologies that acquire raw data from specific data production environments. Subsequently, after collecting raw data, high-speed transmission mechanisms are needed, to transfer the data into the proper storage sustaining system for various types of analytical applications. Finally, collected datasets might contain many meaningless data, which unnecessarily increase the amount of storage space required and adversely affect the consequent data analysis [11]. For example, high redundancy is very common among datasets collected by sensors for environment monitoring. Data compression technology can be applied to reduce the redundancy. Therefore, data pre-processing operations are indispensable to ensure efficient data storage and exploitation [12].

Special data collection techniques are utilized in order to acquire raw data from specific data generation environments. This statement refers to the process of retrieving raw data from real-world objects. The process needs to be well designed [13]. Otherwise, inaccurate data collection would impact the subsequent data analysis procedure and ultimately lead to invalid results. At the same time, data collection methods not only depend on the physical characteristics of the data sources, but also on the objectives of data analysis. As a result, there are many kinds of data collection methods. In the following sections, three common methods for big data collection will be explained, while some related methods will be outlined [14].

Data Collection methods

1. **Log files:** As one widely used data collection method, log files are record files automatically generated by the data source system, so as to record activities in

designated file formats for subsequent analysis. Log files are typically used in nearly all digital devices. For example, web servers record in log files number of clicks, click rates, visits, and other property records of web users [15]. To capture activities of users at the web sites, web servers mainly include the following three log file formats: public log file format (NCSA), expanded log format (W3C), and IIS log format (Microsoft). All the three types of log files are in the ASCII text format. Databases other than text files may sometimes be used to store log information to improve the query efficiency of the massive log store [16, 17]. There are also some other log files based on data collection, including stock indicators in financial applications and determination of operating states in network monitoring and traffic management.

2. **Web Crawlers:** A crawler [18] is a program that downloads and stores webpages for a search engine. Roughly, a crawler starts with an initial set of URLs to visit in a queue. All the URLs to be retrieved are kept and prioritized. From this queue, the crawler gets a URL that has a certain priority, downloads the page, identifies all the URLs in the downloaded page, and adds the new URLs to the queue. This process is repeated until the crawler decides to stop. Web crawlers are general data collection applications for website-based applications, such as web search engines and web caches. The crawling process is determined by several policies, including the selection policy, re-visit policy, politeness policy, and parallelization policy [19]. Traditional web application crawling is a well-researched field with multiple efficient solutions. With the emergence of richer and more advanced web applications, some crawling strategies [20] have been proposed to crawl rich Internet applications. Currently, there are plenty of general-purpose crawlers available as enumerated in the list [21].

3. **Other methods:** In addition to the methods discussed above, there are many data collection methods or systems that pertain to specific domain applications. For example, in certain government sectors, human biometrics [22], such as fingerprints and signatures, are captured and stored for identity authentication and to track criminals.

Data Collection Tools

The role of technology could easily be integrated into various subtopics on crisis and disaster management. The advantages in sensing, networking and communication produce improvements in crisis management from both the research and practice perspectives. Technological advances are necessary to promote the effectiveness of crisis management systems. Reference must be made to the role Geographical Information Systems (GIS), the Global Positioning System (GPS) and Remote Sensing Technologies have in the context of data acquisition [23].

Geographical Information Systems are informative systems capable of storing, analyzing, sharing, and displaying geographically referenced information data. With the usage of GIS crisis administrators are in position to collect spatial information over a wide geographic area, to analyze and collect up to date information. In addition, given the information from GIS can be easily tabulated, providing a pictorial overview of what happening in area

was hit by the crisis. GIS applications can be useful in the following activities:

- To promote situational awareness. Situational awareness is a prerequisite in any
- To create hazard inventory maps. At this level GIS can be used for the pre-feasibility study of developmental projects, at all inter-municipal or district level.
- Locate critical facilities. The GIS system is quite useful in providing information on the physical location of shelters, drains and other physical facilities. The use of GIS for disaster management is intended for planners in the early phase of regional development projects or large engineering projects.
- Create and manage associated databases. The use of GIS at this level is intended for planners to formulate projects at feasibility levels, but it is also used to generate hazard and risk maps for existing settlements and cities, and in the planning of disaster preparedness and disaster relief activities.
- Vulnerability assessment. GIS can provide useful information to boost disaster awareness with government and the public, so that (on a national level) decisions can be taken to establish or expand disaster management organizations. At such a general level, the objective is to give an inventory of disasters and simultaneously identify “high-risk” or vulnerable areas within the country.

GIS technology can provide the user with accurate information on the exact location of an emergency situation. This would prove useful as less time is spent trying to determine where the trouble areas are. Ideally, GIS technology can help to provide quick response to an affected area once issues are known. Mapping and geo-spatial data will provide a comprehensive display on the level of damage or disruption that was sustained as a result of the emergency. GIS can provide a synopsis of what has been damaged, where, and the number of persons or institutions that were affected. This kind of information is quite useful to the recovery process. [24]. An indispensable tool provided by GIS technologies is the GRP, that facilitates real time tracking of the accurate position of parties of interest. By the use of suitable hardware, GPS can be used for a variety of activities from navigation to observing volcanic activity [25].

Remote Sensing

Remote sensing refers to sensors that are attached to aircrafts or satellites. Robotic vision systems the use of remote sensing shows the following features: Data

acquisition far away from the emergency area, regular renewal of the data and also provides big image data of very large areas. [26] Sensors also are used commonly to measure a physical quantity and convert it into a readable digital signal for processing (and possibly storing). Sensor types include acoustic, sound, vibration, automotive, chemical, electric current, weather, pressure, thermal, and proximity. Through wired or wireless networks, this information can be transferred to a data collection point. Wired sensor networks leverage wired networks to connect a collection of sensors and transmit the collected information. This scenario is suitable for applications in which sensors can easily be deployed and managed. For example, many video surveillance systems in industry are currently built using a single Ethernet unshielded twisted pair per digital camera wired to a central location. [27]

Social Media

Big Data analytics provides a great opportunity to reveal many sources of data. Exploring social media represents a significant challenge for big data analytics in crisis and disaster management. Research has emerged that deals with monitoring the trends of social media like Facebook, twitter, etc, before or during times of crisis. Thus, social media represent another big data source of interest. [28]

C. Data storage

The explosive growth of data imposes strict requirements on storage and management. Big data storage refers to the storage and management of large-scale datasets while achieving speed, reliability and availability of data access. It is necessary to review important issues including massive storage systems, distributed storage systems, and big data storage mechanisms. On one hand, the storage infrastructure needs to provide information storage service with reliable storage space; on the other hand, it must provide a powerful access interface for query and analysis of a large amount of data.[29] The data storage subsystem in a big data platform organizes the collected information in a convenient format for analysis and value extraction. For this purpose, the data storage subsystem should provide two sets of features:

1. The storage infrastructure must accommodate information persistently and reliably.
2. The data storage subsystem must provide a scalable access interface to query and analyze a vast quantity of data.

This functional decomposition shows that the data storage subsystem can be divided into hardware infrastructure and data management tools. Hardware infrastructure is responsible for physically storing the collected information. The storage infrastructure can be understood from different perspectives. Typical storage technologies include RAM and cache memory, hard disk drives and disk arrays.

Storage infrastructure can be classified from a networking architecture perspective [30]. In this category, the storage subsystem can be organized in different ways, including, but not limited to the following.

Direct Attached Storage (DAS): DAS is a storage system that consists of a collection of data storage devices. These devices are connected directly to a computer through a host bus adapter (HBA) with no storage network between them and the computer. DAS is a simple storage extension to an existing server.

Storage Area Network (SAN): SANs are dedicated networks that provide block-level storage to a group of computers.

SANs can consolidate several storage devices, such as disks and disk arrays, and make them accessible to computers such that the storage devices appear to be locally attached devices.[31]

Network Attached Storage (NAS): NAS is file-level storage that contains many hard drives arranged into logical, redundant storage containers. Compared with SAN, NAS provides both storage and a file system, and can be considered as a file server, whereas SAN is volume management utilities, through which a computer can acquire disk storage space.

Crisis management data storage tools

Storage mechanisms for big data may be classified into three bottom-up levels: file systems, databases, and programming models. File systems are the foundation of the applications at upper levels. Google's GFS is an expandable distributed file system to support large-scale, distributed, data-intensive applications [32]. GFS uses cheap commodity servers to achieve fault-tolerance and provides customers with high performance services. GFS supports large-scale file applications with more frequent reading than writing. However, GFS also has some limitations, such as a single point of failure and poor performances for small files. Such limitations have been overcome by Colossus [33], the successor of GFS. In addition, other companies and researchers also have their solutions to meet the different demands for storage of big data. For example, HDFS and Kosmosfs are derivatives of open source codes of GFS. Microsoft developed Cosmos [34] to support its search and advertisement business. Facebook utilizes Haystack [35] to store the large amount of small-sized photos. Taobao also developed TFS and FastDFS. In conclusion, distributed file systems have been relatively mature after years of development and business operation. Some of the available tools to facilitate big data storage are:

1. *Google BigTable:* a distributed, structured data storage system, which is designed to process the large-scale (PB class) data among thousands commercial servers [36]. The basic data structure of BigTable is a multi-dimension sequenced mapping with sparse, distributed, and persistent storage. Indexes of mapping are row key, column key, and timestamps, and every value in mapping is an unanalyzed byte array. BigTable is based on many fundamental components of Google, including GFS [37], cluster management system, SSTable file format, and Chubby [38]. GFS is used to store data and log files.
2. *Cassandra:* a distributed storage system to manage the huge amount of structured data distributed among multiple commercial servers [39]. The system was developed by Facebook and became an open source tool in 2008. It adopts the ideas and concepts of both Amazon Dynamo and Google BigTable, especially integrating the distributed system technology of Dynamo with the BigTable data model. Tables in Cassandra are

in the form of distributed four-dimensional structured mapping, where the four dimensions including row, column, column family, and super column. The partition and copy mechanisms of Cassandra are very similar to those of Dynamo, so as to achieve consistency.[40]

3. *Hadoop:* a top level Apache project that started in 2006. Hadoop can process extremely large volume of data with different structures. Is used commonly in industrial applications, analyzes big data with specific functions such as spam filtering, network click stream analysis and social recommendations. [41, 42]. In fact, Hadoop has long been the mainstay of the big data movement, Instead of relying on expensive, proprietary hardware to store and process data, Hadoop enables distributed processing of large amounts of data on large clusters of commodity servers. Hadoop offers scalability, cost efficiency, flexibility and fault tolerance. Hadoop can recover the data and computation failures caused by node breakdown or network congestion. The Apache Hadoop software library is a massive computing framework consisting of several modules, including HDFS, Hadoop MapReduce, HBase, and Chukwa. [43]
4. *MapReduce:* a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The computational model consists of two user defined functions, called Map and Reduce. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks. [44] The concise MapReduce framework only provides two opaque functions, without some of the most common operations (e.g. Projection and filtering). [45]
5. *Dryad:* a general-purpose distributed execution engine for processing parallel applications of coarse-grained data. The operational structure of Dryad is a directed acyclic graph, in which vertices represent programs and edges represent data channels. Dryad executes operations on the vertices in clusters and transmits data via data channels, including documents, TCP connections, and shared-memory FIFO. All kinds of data are directly transmitted between vertexes [46]. In addition, Dryad allows vertexes to use any amount of input and output data, while MapReduce supports only one input and output set. DryadLINQ [47] is the advanced language of Dryad and is used to integrate the aforementioned SQL-like language execution environment
6. NOSQL databases (non – relational databases)

With the development of the Internet and cloud

computing, there need databases to be able to store and process big data effectively, demand for high-performance when reading and writing, so the traditional relational database is facing many new challenges.[48] Various database systems are developed to handle datasets at different scales and support various applications. Traditional relational databases cannot meet the challenges on categories and scales brought about by big data. NoSQL databases (i.e., non traditional relational databases) are becoming more popular for big data storage. [49] Especially in large scale and high-concurrency applications, such as search engines and SNS, using the relational database to store and query dynamic user data has appeared to be inadequate. [50]

D. Data Analysis

The last and most important stage of the big data value chain is data analysis, the goal of which is to extract useful values, suggest conclusions and/or support decision-making. Firstly, the purpose and classification metric of data analytics will be discussed. Subsequently, the application evolution for various data sources and summarize the six most relevant areas will be reviewed. Finally, several common methods that play fundamental roles in data analytics will be introduced. Data analytics addresses information obtained through observation, measurement, or experiments about a phenomenon of interest. The aim of data analytics is to extract as much information as possible that is pertinent to the subject under consideration. The nature of the subject and the purpose may vary greatly. Some example aims include:

- To extrapolate and interpret the data and determine how to use it,
- To check whether the data are legitimate,
- To give advice and assist decision-making,
- To diagnose and infer reasons for fault, and
- To predict what will occur in the future

In [53] data analytics are classified into three levels according to the depth of analysis: descriptive analytics, predictive analytics, and prescriptive analytics.

Descriptive Analytics: exploits historical data to describe what occurred. For instance, a regression may be used to find simple trends in the datasets, visualization presents data in a meaningful fashion, and data modeling is used to collect, store and cut the data in an efficient way. Descriptive analytics is typically associated with business intelligence or visibility systems.

Predictive Analytics: focuses on predicting future probabilities and trends. For example, predictive modeling uses statistical techniques such as linear and logistic regression to understand trends and predict future outcomes, and data mining extracts patterns to provide insight and forecasts.

Prescriptive Analytics: addresses decision making and efficiency. For example, simulation is used to analyze complex systems to gain insight into system behavior and identify issues and optimization techniques are used to find optimal solutions under given constraints. [54]

III. BIG DATA IN CRISIS PHASES

Crisis

Professor C. Hermann in his article in Administrative Science magazine in June 1963 [55] states that “the crisis is a condition characterized by surprise, a high risk of serious values and short reaction time”. The four phases of crisis are: Prevention, Preparedness, Response and Recovery. These formulate the crisis cycle. There are many interesting approaches about the usage of Big Data in crisis management.

Big Data and Crisis Prevention

Information derived from the analysis of Big Data can help to anticipate crises or at least reduce the risks that would arise from a disaster the major crisis effect. One example is in a big earthquake harm arises in telecommunication networks leading to interruption of communications, also has been observed a large number of blackouts. There exists a need to study this data for optimization of the civil infrastructure to avoid this crisis effects. [56]

Big Data and Crisis Preparedness

Big Data analysis can help significantly to the preparation of crisis management. Through the data analysis can be done recognizing the dangers and to provide a sound strategic approach by the respective managers of the crisis. Big Data analysis can also guide the proactive deployment of resources to fully cope with an impending type of disaster [57]

Big Data and Crisis Response

Big Data analysis in real time can identify which areas need the most urgent attention from the crisis administrators. With the use of the GIS and GPS systems, Big Data analysis can assist the right guidance to the public to avoid or move away from the hazardous situation. Furthermore analysis from prior crisis could help identify the most effective strategy for responding to future disasters. [58]

Big Data and Crisis Recovery

When the recovery activation will gradually start, the infrastructure would provide a big data source. The Big Data analysis sharing useful information for recovery procedures about volunteer coordination and logistics during the crisis. [59]

IV. CONCLUSION

In this paper, the usefulness of the analysis of Big Data management in crises and disasters was presented. A brief analysis of the collection data sources during the crisis, the technological means and the tool storage and processing of Big Data. The challenges arising from the review concerns the important research fields of the Social media data usage in crisis management. In this context, a system-engineering approach of a big data management system into four phases, data generation, data acquisition, data storage, and data analytics was also outlined. The era of big data is upon us, bringing with it an urgent need for advanced data acquisition, management, and analysis mechanisms. In the big data acquisition phase, typical data collection technologies were investigated during each stage of the data

life cycle the management of big data is the most demanding issue. Many challenges in the big data system need further research attention. Big data research remains in its embryonic period. Research on typical big data applications, is required that can improve the efficiency of government sectors, and promote the development of human science and technology, while it is also required to accelerate big data progress. Furthermore there are interesting challenges in data mining in crisis and disasters management. Algorithms need to be developed for completing tasks such as pattern mining for discovering interesting associations and correlations, clustering and trend analysis, to understand the nontrivial changes and trends, and classification to prevent future reoccurrences of undesirable phenomena. Finally several security challenges in storage and transmission of data need to be under constant investigation, in order to address newly emerging threats.

REFERENCES

- [1] S. Kailser, F. Armour, J. A. Espinosa and W. Money, "Big Data: Issues and Challenges Moving Forward," 46th Int. Conf. System Sciences, pp. 995,
- [2] S. Kailser, F. Armour, J. A. Espinosa and W. Money, "Big Data: Issues and Challenges Moving Forward," 46th Int. Conf. System Sciences, pp. 996-997,
- [3] S. Kailser, F. Armour, J. A. Espinosa and W. Money, "Big Data: Issues and Challenges Moving Forward," 46th Int. Conf. System Sciences, pp. 996-997,
- [4] S. Mehrotra, X. Qiu, Z. Cao, and A. Tate, "Technological Challenges in Emergency Response", "(Periodical style)," IEEE, pp. 6 July/August 2013 <https://www.computer.org/intelligent>.
- [5] S. Mehrotra, X. Qiu, Z. Cao, and A. Tate, "Technological Challenges in Emergency Response", "(Periodical style)," IEEE, pp. 6 July/August 2013 <https://www.computer.org/intelligent>
- [6] F. Gallagher. (2013). The Big Data Value Chain [Online]. Available: <http://fraysen.blogspot.sg/2012/06/big-data-value-chain.html>
- [7] E. B. S. D. D. Agrawal et al., "Challenges and opportunities with big Data" A community white paper developed by leading researchers across the united states," The Computing Research Association, CRA White Paper, Feb. 2012.
- [8] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, "Interactions with big data analytics," Interactions, vol. 19, no. 3, pp. 50-59, May 2012.
- [9] What is Big Data, IBM, New York, NY, USA [Online]. Available: <http://www-01.ibm.com/software/data/bigdata/> 2013
- [10] J. Manyika et al., Big data: The Next Frontier for Innovation, Competition, and Productivity. San Francisco, CA, USA: McKinsey Global Institute, 2011, pp. 1-137.
- [11] H. Hu, Y. Wen, T. Chua and X. Li, "Toward Scalable System for Big Data Analytics: a Technology Tutorial," (Periodical style)," IEEE pp 8 July 2014 657-659.
- [12] M. Chen, S. Mao, and Y. Liu , " Big data : a survey", Mobile Networks and Applications, vol.19, no.2, pp.181-183, 2014
- [13] H. Hu, Y. Wen, T. Chua and X. Li, "Toward Scalable System for Big Data Analytics: a Technology Tutorial," (Periodical style)," IEEE pp 8 July 2014 659-663.
- [14] M. Chen, S. Mao, and Y. Liu , " Big data : a survey", Mobile Networks and Applications, vol.19, no.2, pp.181-183, 2014
- [15] Wahab MHA, Mohd MNH, Hanafi HF, Mohsin MFM (2008) Data pre-processing on web server logs for generalized association rules mining algorithm. World Acad Sci Eng Technol 48:2008
- [16] A. Nanopoulos , Y. Manolopoulos , M. Zakrzewicz and T. Morzy , (2002) Indexing web access-logs for pattern queries. In: Proceedings of the 4th international workshop on web information and data management. ACM, pp 63-68
- [17] K. Joshi , A. Joshi ,Y. Yesha, (2003) On using a warehouse to analyze web logs. Distrib Parallel Databases 13(2):161-180
- [18] J. Cho and H. Garcia-Molina, "Parallel crawlers," in Proc. 11th Int. Conf. World Wide Web, 2002, pp. 124-135
- [19] C. Castillo, "Effective web crawling," ACM SIGIR Forum, vol. 39, no.1, pp. 55-56, 2005.
- [20] S. Choudhary et al., "Crawling rich internet applications: The state of the art." in Proc. Conf. Center Adv. Studies Collaborative Res.(CASCON), 2012, pp. 146-160.
- [21] (2013, Oct. 31). Robots [Online]. Available: <http://user-agentstring.info/list-of-ua/bots>
- [22] A. K. Jain, et al., Biometrics: Personal Identification in Networked Society. Norwell, MA, USA: Kluwer, 1999.
- [23] Introduction to Disaster Management, Virtual University for Small States of the Commonwealth (VUSSC) pp 97-129, July 2011
- [24] Introduction to Disaster Management, Virtual University for Small States of the Commonwealth (VUSSC) pp 97-129 July 2011
- [25] Introduction to Disaster Management, Virtual University for Small States of the Commonwealth (VUSSC) pp 97-129 July 2011
- [26] V. Hristidis, S. Chen, T. Li, S. Luis and Y. Deng, "Survey of Data Management and Analysis in Disaster Situations. (Periodical style)," ELSEVIER June 2010 <https://www.elsevier.com/locate/jss>
- [27] M. Chen, S. Mao, and Y. Liu , " Big data : a survey", Mobile Networks and Applications, vol.19, no.2, pp.171-209, 2014.
- [28] S.Chaudhuri, "What Next? A Half-Dozen Data Management Research Goals for Big Data and the Cloud"(Periodical style)," Proceedings of the 1st symposium on Principles of Database Systems, ACM, 2012.
- [29] M. Chen, S. Mao, and Y. Liu , " Big data : a survey", Mobile Networks and Applications, vol.19, no.2, pp.184-185, 2014.
- [30] U. Troppens, R. Erkens, W. Mueller-Friedt, R. Wolafka, and N. Haustein, Storage Networks Explained: Basics and Application of Fibre Channel SAN, NAS, ISCSI, FCoE. New York, NY, USA: Wiley, 2011.
- [31] H. Hu, Y. Wen, T. Chua and X. Li, "Toward Scalable System for Big Data Analytics: a Technology Tutorial," (Periodical style)," IEEE pp 8 July 2014 665-666.
- [32] R. Cattell Scalable sql and nosql data stores. ACM SIGMOD Record 39(4):12-27, 2011
- [33] McKusick MK, Quinlan S. "Gfs: evolution on fastforward". ACM Queue 7(7):10, 2009
- [34] R. Chaiken, B. Jenkins, Larson, P-A°. Ramsey B, D. Shakib, S. Weaver, J. Zhou Scope: "easy and efficient parallel processing of massive data sets. Proc VLDB Endowment "1(2):1265-1276, 2008
- [35] D. Beaver, S. Kumar, Li HC, J. Sobel, P. Vajgel et al (2010) Finding a needle in haystack: facebook's photo storage. In OSDI, vol 10. pp 1-8
- [36] Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE (2008) Bigtable: a distributed storage system for structured data. ACM Trans Comput Syst (TOCS) 26(2):4
- [37] R. Cattell (2011) Scalable sql and nosql data stores. ACM SIGMOD Record 39(4):12-27
- [38] M. Burrows (2006) The chubby lock service for loosely-coupled distributed systems. In: Proceedings of the 7th symposium on Operating systems design and implementation. USENIX Association, pp 335-350
- [39] Lakshman A, Malik P (2009) Cassandra: structured storage system on a p2p network. In: Proceedings of the 28th ACM symposium on principles of distributed computing. ACM, pp 5-5
- [40] M. Chen, S. Mao, and Y. Liu , " Big data : a survey", Mobile Networks and Applications, vol.19, no.2, pp.187-190, 2014.
- [41] S. Sagirolglou and D. Sinanc, "Big data : a review," in Proceedings of the International Conference on Collaboration Technologies and Systems (CTS'13), pp 42-47, IEEE, San Diego, Calif, USA, May 2013.
- [42] S. Sagirolglou and D. Sinanc, "Big data : a review," in Proceedings of the International Conference on Collaboration Technologies and Systems (CTS'13), pp 42-47, IEEE, San Diego, Calif, USA, May 2013.
- [43] J. H. Howard et al., "Scale and performance in a distributed le system," ACM Trans. Comput. Syst., vol. 6, no. 1, pp. 51-81, 1988.
- [44] D. Laney (2001) 3-d data management: controlling data volume, velocity and variety. META Group Research Note, 6 February
- [45] P. Zikopoulos, C. Eaton., (2011) Understanding big data: analytics for enterprise class hadoop and streaming data. McGraw-Hill
- [46] Isard M, Budiu M, Yu Y, Birrell A, Fetterly D. Dryad: distributed data-parallel programs from sequential building blocks. ACM SIGOPS Oper Syst Rev 41(3):59-72. 2007
- [47] Yu Y, Isard M, Fetterly D, Budiu M, Erlingsson U , Gunda PK, Currey Dryadling: a system for general-purpose distributed data-parallel computing using a high-level language. In: OSDI, vol 8. pp 1-14, 2008.
- [48] J. Han, E. Haihong, G. Lee, J. Du. "Survey on NoSQL database" Pervasive Computing and Applications (ICPCA), 2011 6th International Conference. Pp 363-366, IEEE, 26-28 Oct. 2011
- [49] M. Chen, S. Mao, and Y. Liu , " Big data : a survey", Mobile Networks and Applications, vol.19, no.2, pp.186, 2014

- [50] J.Han, E. Haihong, G.lee, J.Du. "Survey on NoSQL database" Pervasive Computing and Applications (ICPCA), 2011 6th International Conference. Pp 363-366, IEEE, 26-28 Oct. 2011
- [51] Karger D, Lehman E, Leighton T, Panigrahy R, Levine M, Lewin D (1997) Consistent hashing and random trees: distributed caching protocols for relieving hot spots on the world wide web. In: Proceedings of the twenty-ninth annual ACM symposium on theory of computing. ACM, pp 654–663.
- [52] M. Chen, S. Mao, and Y. Liu , " Big data : a survey", Mobile Networks and Applications, vol.19, no.2, pp.186, 2014
- [53] G. Blackett. Analytics Network-O.R. Analytics [Online]. Available:http://www.theorsociety.com/Pages/SpecialInterest/AnalyticsNetwork_analytics.aspx, 2013.
- [54] H. Hu, et al., "Toward Scalable System for Big Data Analytics: a Technology Tutorial," IEEE pp 671-672. July 2014
- [55] C. Hermann," Some Consequences of Crisis which Limit the Viability of Organizations" Administrative Science Quarterly (pp 61-82), 1963.
Big Data and Disaster Management, JST/NSF joint workshop, pp, 7, 8, 20 July 2011

Identifying Peer-to-Peer Traffic Based on Traffic Characteristics

Prof S. R. Patil

Dept. of Computer Engineering
SIT, Savitribai Phule Pune University
Lonavala, India
srp.sit@sinhgad.edu

Suraj Sanjay Dangat

Dept. of Computer Engineering
SIT, Savitribai Phule Pune University
Lonavala, India
surajdangat6691@gmail.com

Abstract—The P2P (Peer-to-Peer) network is dynamic, self-organized and has some other features. So, P2P traffic has become one of the most significant portions of the network traffic. But, it has also caused network congestion problems because of resource occupation (mainly bandwidth). Accurate identification of P2P traffic makes great sense for efficient network management and efficient utility of network resources. In this paper we address traffic identification technology, such as traffic identification by using network characteristics have been considered for solving the problem with different parameters, improved accuracy rate and efficiency. Experiments on various P2P applications demonstrate that the method is generic and it can be applied to most of P2P applications. Experimental result shows that the algorithm can identify P2P application accurately. In this paper we first briefly introduce P2P technology and then we made a short survey on the overall progress in P2P traffic identification technologies. Finally we discuss the proposed method and its result analysis.

Keywords—P2P; Peer-to-Peer, Traffic characteristics, Peer-to-Peer Traffic identification

I. INTRODUCTION

A network is a group or system of interconnected people or some things. A computer network or data network is a telecommunications network in which a group of two or more computer systems linked together and it allows computers to exchange data. In computer networks, networked computing devices pass data to each other along network links and these links between nodes are established using either wired media or wireless media. The widely-known example of computer network is the Internet. In a P2P network, the "peers" are computer systems which are connected to each other via the Internet and peers are configured to allow certain files and folders to be shared with every peer or with selected peers. Files can be shared directly between the different peers on the network without the need of a central server. In short, each computer on a P2P network acts as a file server as well as a client.

With the extensive application of P2P technology, P2P applications consume a large amount of network bandwidth, which ultimately increase the burden of the network. According to the available statistics, P2P applications account for 60% to 80% of total ISP business and become the largest consumer of network bandwidth. According to the statistics,

about 60 percent of the bandwidth is occupied by P2P applications and from these 60%, 80% of which were occupied by P2P file-sharing applications. But these P2P file-sharing users are very low in number and they are only 5% of the total number of Internet users. The P2P has many characterizes, such as large flow, automatic operation, connection for a long time, regardless of time running, and so on. Therefore, P2P applications can take up more bandwidth than other applications. As number of P2P user's increases, network traffic will also increase significantly. At the same time, increasing the size of the network will lead a large number of broadcast news to flood in the entire network and so the network traffic increases. In the end, it led to bottlenecks of the network and network congestion damages the services provided by the Internet Service Providers and common users [1].

The solution to this bandwidth congestion problem is that we limit the users who use large amount of bandwidth to protect those user who use a small amount of bandwidth when the network resource constraints. On the contrary, when there are no constraints on network resources, we remove these restrictions so that each user can use the lines efficiently. How to effectively control the available network resources and how to effectively control the P2P traffic have become quite important questions. Therefore, P2P traffic identification is a key technology for effective control over it.

P2P traffic which is produced by P2P application has different characteristics than the other applications. In P2P traffic identification based on traffic characteristics, these characteristics are used to identify the network traffic. This approach can detect P2P traffics with dynamic ports and encrypted transmission. This method easily detects the flow of encrypted payload and unknown payload characteristics of P2P traffic [2].

In recent years, researchers have been using traffic characteristics in the area of computer network for identifying P2P traffic. It is very relevant for the study of network traffic control, traffic congestion, resource utilization and it also has an economic importance. In this paper, the main focus is on three things i.e. on classifying all the incoming traffic in the network, identify the P2P traffic in the network and to minimize the rate of false detection and the rate of negative detection. The rest sections of this paper are arranged as

follow. In section II, we briefly introduce Details of P2P traffic classification technologies. In Section III, we briefly discuss the System Architecture. In Section IV, we represent proposed method by means of Mathematical model. In Section V, we analyse the Result of proposed system. In Section VI, we discuss the conclusion of proposed system.

II. EXISTING METHODS OF P2P TRAFFIC CLASSIFICATION TECHNOLOGIES

There are various techniques available for P2P traffic identification and classification. It includes Port-based classification, Payload-based classification, Feature-based classification and Hybrid classification method. It is discussed in following paragraphs.

Port-based classification method is the simplest and traditional method. It identifies the application traffic by identifying the application type and this application type is identified from the port number used in the transport layer. For example, TCP port 80/443/1024 is Skype traffic, TCP port 1214 is Kazaa P2P traffic and so on. This approach is extremely easy for implementation and it gives very little overhead on the traffic classifier. But, it also has some limitation and it becomes less accurate because of several reasons. These are, many applications uses random ports and some P2P applications use dynamic ports which are not known in advance [3].

To remove the drawbacks of port-based classification method, several payload-based classification techniques have been proposed. Most protocols contain a protocol specific string in the payload namely signatures that can be used for identification. The information about this strings are publicly available. Subhabrata [4] presented an analysis of a number of P2P application and their signatures. For example "0x13Bit" corresponds to the BitTorrent application. By comparing every packet payload with a set of previously determined signatures, this method can identify application traffic more accurately than the traditional Port-based method. The benefits of this method are high accuracy and robustness, and have a good classification functions. However, there are still some disadvantages of this method. These methods identify only P2P traffic for those signatures which is known in advance and it is unable to classify any other traffic.

Because of the disadvantages of these two methods, the research community started developing the new methods which are less dependent on particular individual applications, but focused on capturing and extracting common things in the behavior of P2P applications. We refer this as feature-based technique. This kind of approach is to classify traffic based on the analysis of hidden transition patterns of traffic flows. Such nonlinear properties cannot be affected by dynamic port change or payload encryption. These methods provide an alternative for effective traffic classification.

There are also some hybrid P2P traffic classification methods available. This approach includes most of the proposed methods for improving classification accuracy. Most of the traffic identification method of P2P based on traffic characteristics is to select particular feature from the P2P features [5, 6]. There are four main P2P features. First is that

the number of other hosts that P2P hosts are connecting to be bigger than the traditional hosts. Second, P2P traffics of up and down are roughly equal and it is different from the traditional characteristics of the host. Third, P2P hosts are act as a both servers as well as clients, which differ from the traditional hosts. Fourth, the connection features of P2P hosts listening port are different from the traditional ones [7]. In the reference [8], the authors proposed a simple algorithm based on the first feature. In the second reference, the authors proposed a flow characteristics identification algorithm based on first and fourth feature and also uses factors that affect the efficiency of identification algorithm. Proposed system considers different features and uses K-means and Naive-Bayes algorithm to improve accuracy rate. It minimizes the rate of false detection and the rate of the negative detection. The next section describes the system architecture of the proposed system.

III. SYSTEM ARCHITECTURE

The system architecture of our proposed system is discussed briefly below. It is shown in following figure1. It has two phases, In Phase 1 we calculate the optimum thresholds and in Phase 2 we apply these optimum thresholds values for classifying P2P traffic. Input to the proposed system is all traffic. We get output as classified traffic. There are one preprocessing module and three main modules. In preprocessing module, means in Calculate Optimum Threshold We calculate the optimum threshold value in it. Main modules are namely Tracking Traffic Model, Traffic Identification Module and Traffic Classification.

A. Module I-Tracking Traffic Model

In this module, System tracks the traffic using Packet Sniffer and represents it in the form of array of packet objects. Input to this module is all traffic. The output of this module is Tracked Traffic. There are two steps in it and these are Packet Sniffer and Incoming Packets Dataset. In Packet Sniffer step all traffic is passed to the Packet Sniffer and it continually tracks the maximum traffic as much as possible. Then in Incoming Packets Dataset step, all traffic which is tracked by Packet Sniffer is representing in the form of Packet class objects. The array for this is maintained.

B. Module II-Traffic Identification Module

In this module three things take place and these are packet analysis, feature set, packet classification takes places. Input to this module is the array of packets. Output of this module is classified packets. There are three steps in it and these are Packet Analyzer, Feature Set and Packet Classification. In Packet Analyzer step, the array of packets which is tracked by Packet Sniffer is passed to Packet Analyzer to analyze each packet in details. Additional information which is getting by this step we represent and maintained using collection objects.

In Feature Set step, system maintains the particular features and its ideal value for packet classification. These features include packets destination IP, hop limit, packet length and port number. We determine the optimal threshold value for particular traffic in preprocessing module namely calculate

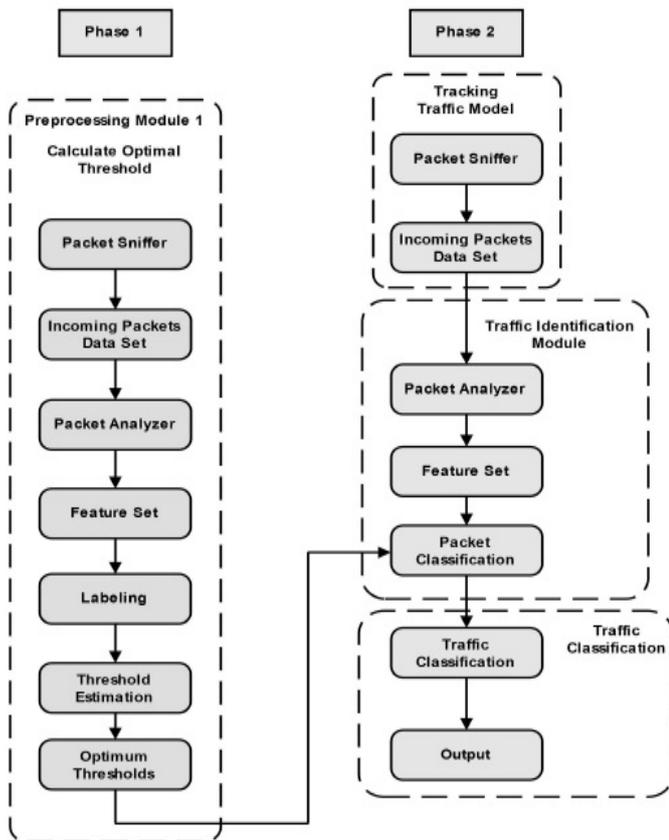


Fig. 1. State System architecture of proposed system

Optimum Thresholds and it is part of it. We apply K-Means algorithm in it. First we consider trainee features of packet then we find two centroids for packet length that is Centroid 1 for minimum value and Centroid 2 for its maximum value. K-means is one of the simplest unsupervised learning algorithms that solve the widely-known clustering problem. The procedure of K-means algorithm follows a simple and it is easy way to classify a given data set through a certain number of clusters assume as a k clusters, fixed a priori. The main idea is to define k centroids, one for each cluster and these centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other and then, the next step is to take each point belonging to a given data set and associate it to the nearest centroid.

In Packet Classification step all packets then classified by using Naive-Bayes classifier algorithm. The input to this step is calculated optimum threshold values, collection of feature set and tracked packet array. We consider the min and max values of packet length of packets in it which is calculated in preprocessing module. We also consider hop limit and port number features. The naive Bayes algorithm is a simple probabilistic classification algorithm. In simple terms, a naive Bayes classifier assumes that the presence or absence of a particular feature of a class is not related to the presence or absence of any other feature, given the class variable. For example, a fruit may be considered to be a tomato if it is red, round, and about 3" in diameter. Even if these features depend

on each other or upon the existence of the other features a naive Bayes classification algorithm considers all of these properties to independently contribute to the probability that this fruit is a tomato [9]. Same thing is applied for each packet. The naive Bayes consider all the properties of packet and its value to independently contribute to the probability that this packet is P2P packet or Non-P2P packet.

C. Module III-Traffic Classification

In this, system classifies the traffic and represents the output to user in both textual and graphical form. Input to this module is classified packets. Output of this module is classified traffic. There are two steps in it and these are Traffic Classification and Output. In Traffic Classification step we classify the particular traffic. In Output step system represents the output as two sets namely as P2P and Non-P2P traffic. This step also represents the output in graphical form.

IV. MATHEMATICAL MODEL

D. System description by means of mathematical formulas

As Let, S is a system which is defined in following manner:

$$S = \{P, PS, PA, FE, MD, KM, C, CL, NB, O, G, F|f_1, f_2, f_3, f_4, f_5, f_6, f_7\}$$

Where,

P = Set of Packets

PS = Packet Sniffer

PA = Set of Analyzed Packets

FE = Feature Extraction

MD = Manage Database

KM = K-Means Algorithm

C = Set of Centroids

CL = Set of Clusters

NB = Nave-Bayes Algorithm

O = Set of Output

G = Set of Graphs

SET THEORY:

$$P = \{P_0, P_1, \dots, P_n\}$$

$$PA = \{PA_0, PA_1, \dots, PA_n\}$$

$$C = \{C_0, C_1, \dots, C_n\}$$

$$CL = \{CL_0, CL_1, \dots, CL_n\}$$

$$O = \{P2P, Non - P2P\}$$

$$G = \{G_0, G_1, \dots, G_n\}$$

TABLE I. MAPPING TABLE

Sr. No.	Function	Description
1	f1(P) → P	Tracking traffic by using Packet Sniffer.
2	f2(P) → PA	Analyze Packet using Packet Analyzer.
3	f3(PA) → D	For Updating Dataset of the system.
4	f4(D, FE) → PE	Packet of Particular Feature.
5	f5(PE, C, CL) → CL	To Find cluster.
6	f6(CL) → C	To get Classified Traffic.
7	f7(C) → O, G	To represent Output as Set and Graph.

INPUT:

$$P = \{P_0, P_1, \dots, P_n\}$$

OUTPUT:

$$O = \{P2P, \text{Non} - P2P\}$$

$$G = \{G_0, G_1, \dots, G_n\}$$

FUNCTIONS:

f1 = For Packet Sniffer to capture network traffic

f2 = For Packet Analyzing

f3 = For Updating Dataset

f4 = For Feature Extraction

f5 = For Finding Clusters

f6 = For getting Classified Traffic

f7 = Output as Graph

E. Function mapping table

Above functions can be mapped onto the elements of the set. It is shown in table I.

F. State Transaction Diagram

The basic idea of this diagram is to define a machine that has finite states. It represents the states of the system. Following figure represents the state transaction diagram of the proposed system.

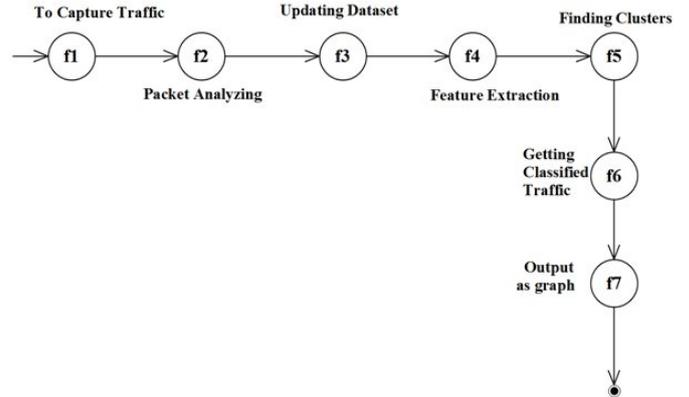


Fig. 2. State transition diagram of system

V. ANALYSIS OF RESULT AND ALGORITHM PERFORMANCE

In campus network it is hard to determine the P2P host. Even we could determine the nodes which use P2P but it is hard to determine how much of traffic is generated by P2P and non-P2P applications. So we conduct the experiments in the laboratory. In that experiment we used some sets of experimental machines which is used for generating enough amount of P2P traffic. It is used for testing robustness and detection accuracy rate of the earlier and proposed methods. These experimental machines generate different types of P2P and Non-P2P traffic during different period of time. We set threshold value and we analyzed network for 1 week. Results of the experiments are shown in Table II.

As we are using K-means for determining min and max threshold values of properties of packet, it makes our analysis and proposed scheme more accurate. The time complexity of proposed method is less than the previous method as we are using Naïve-Bayes classification algorithm. Although the proposed method in this paper still having some weaknesses and false detection rate but they are acceptable.

TABLE II. EFFRCTIVENESS COMPARISON BETWEEN EARLIER TRAFFIC CHARACTERISTICS BASED METHOD AND METHOD IN THE PAPER

Application	Earlier traffic characteristics based method negative rate	Earlier traffic characteristics based method false rate	Expected result of negative rate in the paper	Expected result of false rate in the paper
BT	5.34%	2.86%	2.3%	1.86%
eMule	5.82%	3.23%	2.82%	1.23%
Pplive	4.27%	4.26%	2.29%	2.63%
Kugoo	4.63%	3.25%	2.56%	2.14%
Non-P2P	4.26%	4.84%	2.18%	2.44%

VI. CONCLUSION

The traffic identification is very relevant for the study of network traffic control, traffic congestion, resource utilization and it also has an economic importance. With the P2P technology continues to progress, P2P client has experienced the phase of using fixed-port, random port, encrypted message and the tunnel mode. It is hard for people to identify packets of P2P protocols. Some of the traditional identification

methods like the Port-based identification method, Payload-based identification method and Feature-based identification are not very effective. Traffic identification technologies such as traffic identification by using network characteristics have been considered for solving the problem with improved accuracy rate and improved algorithm efficiency. The previous work of P2P traffic identification based on traffic characteristics, it gives the high rate of false detection and high rate of negative detection and it is not efficient when we consider time as a factor. So we designed a system in such a way that it minimizes the rate of false detection and the rate of negative detection and classifies the traffic in efficient manner. The presented method in this paper can able to detect P2P traffics with dynamic ports and encrypted transmission, and the efficiency of the implementation of the algorithm is also improved to some extent. These techniques are useful for increasing the performance of the network. This system can easily extend for working with other regions of the network in order to solve the problems in it.

ACKNOWLEDGMENT

We are thankful to Mr. S. D. Baber and Mr. P. J. Pandit, for the encouragement and the support they have extended to us for completing this paper. We are also thankful to the Mrs. S. R. Patil and Mr. Sanjay Dangat for their support to make this paper as good as it is. We are also thankful to our family members and friends for patience and encouragement.

REFERENCES

- [1] Yu-shui Geng, Tao Han and Xue-song Jiang "The research of P2P traffic identification technology." (978-1-4244-4589-9/09, 2009, IEEE).
- [2] Jingyu Wang, Jiyuan Zhang, Yuesheng Tan "Research of P2P Traffic Identification Based on Traffic Characteristics"(978-1-61284-774-0/11, 2011, IEEE).
- [3] Jian Feng "Research on the Technology of Peer-to-Peer Traffic Classification."(978-1-4244-5567-6/10, 2010, International Symposium on Computer, Communication, Control and Automation, IEEE).
- [4] S. Subhabrata, S. Oliver, D. Wang, "Accurate, scalable in-network identification of p2p traffic using application signatures," Proc. The 13th international conference on World Wide Web, ACM Press, Oct. 2004, pp. 512-521, doi: 10.1145/988672.988742.
- [5] Ke Xu, Ming Zhang, Mingjiang Ye, Dah Ming Chiu, Jianping Wu "Identify P2P traffic by inspecting data transfer behavior." (0140-3664, Elsevier B.V., 2010, Computer Communications).
- [6] CHENG Wei-qing, GONG Jian, DING Wei "Identifying file-sharing P2P traffic based on traffic characteristics." (15(4): 112120, December 2008, paper number:10058885, Sciencedirect).
- [7] Haiming Jiang, Jianying Zhang, Qingqing Wang, "P2P traffic detection and analysis", J. Computer Technology and Development, 2008, 18 (7): 116-119.
- [8] Chao Wen, Xuefeng ZHENG, "The study of P2P protocol identification method based on the traffic analysis" J. Micro Computer Applications, 2007 (7): 714-717
- [9] Mrs.G.Subbalakshmi, Mr. K. Ramesh, Mr. M. Chinna Rao "Decision Support in Heart Disease Prediction System using Naive Bayes" (ISSN : 0976-5166, Vol. 2 No. 2 Apr-May 2011, Indian Journal of Computer Science and Engineering (IJCSE)).

Influence of IT on Micro Enterprises to pursue Strategic Growth

Satya, Shah, and Syed, Hassan

Abstract— This paper aims to provide an assessment of Information Technology and to identify its influences on microenterprises within UK. It develops an understanding towards providing an evaluative measure against strategic growth within microenterprises. The author's draws upon research questions that derive towards identifying if Microenterprises should be regarded in their own right away from the commonly used term SME (Small-medium Enterprise), and to addressing the perceptions of IT on strategic growth within Microenterprises. The paper also aims to identify whether microenterprises can achieve greater time-efficiencies using IT to pursue strategic growth. To address the research areas the paper examines literature studies within the areas of microenterprises; the strategic growth and influence of information technology (IT) within microenterprises. The research carried out observational case study within a project management microenterprise addressing the general processes, perceptions and challenges addressed in the working environment highlighting particular attention towards administrative deficiency with use of IT and reliability issues with their in-house server and attempting to overcome these solutions. The result from the literature and the observations of the case provides an opportunity to propose towards a development framework to assist microenterprises in achieving strategic growth with the use of IT and overcoming their addressed barriers and challenge.

Keywords—Information Technology, Microenterprises, Strategic Growth, SME.

I. INTRODUCTION

Information Technology Consultancy is a fascinating position to be in the world of Business and Technology; it can help other people and businesses on any scale to identify problems and hopefully provide a viable solution to their needs. It could make daily tasks more efficient, it could result in scrupulous tasks being automated to enable skills to be put to better use. It could save time, save money and generate more business. Large corporate companies have the ability to invest millions into finding ways to innovate, adapt and more

so in today's climate; reduce costs, speed up processes and/or make the company more profit [1]. Teams of experienced professionals with predefined roles will take a problem and design, create and develop a solution and implement it, not to mention at sometimes astronomical costs. But for the 4.6million microenterprises trading in the UK as of 2012, the ability to develop better ways to conduct daily processes and with the prospect of wanting to grow, it can prove difficult with limited time and resources available [2].

Strikingly the number of microenterprises registered has risen by 183% within 5 years, largely down to the economic conditions currently persisting in the UK with people's changing circumstances of increasing unemployment and the ability for Microenterprises to capitalize on their agility and service offering to compete with larger companies [3]. So with the ever evolving technology out there, how is technology helping to leverage microenterprises to better prosperity? How does IT play in the role of enabling a micro enterprise to achieve growth strategically? There is extensive literature around the corporate world of business and strategy and IT evolution, however from initial research it has become apparent very little research available on the influences of IT on SMEs and scarcer when narrowed to just micro enterprises [1]. The research aims to address some of research questions as:

- a) Should microenterprises have greater awareness in society from Small-Medium Enterprises (SMEs)
- b) Are all microenterprises driven to succeed by becoming larger businesses or they look to succeed by remaining 'Micro'
- c) What are the current perceptions of IT and its impact on microenterprises aiming to pursue their current strategy

The authors aims to analyse current literature to identify the importance of classifying microenterprise to their larger counterparts and towards determining factors that influence microenterprise strategic growth with particular focus towards the application of information technology. An observational study was carried out to identify general influences and challenges of information technology within a Microenterprise with the use of semi structured questionnaires. The synthesis of literature studies and that of the observational results will further allow the research to propose a development framework of solutions that could enable strategic growth within microenterprises.

Satya Shah is the Director for Technology Management Programmes in the Faculty of Engineering and Sciences, University of Greenwich, UK (email: s.shah@gre.ac.uk)

Syed Hasan is Assistant Professor in the Department of Industrial & Manufacturing Engineering at NED University of Engineering and Technology, Pakistan (email: syedhasan@neduet.edu.pk)

II. LITERATURE REVIEW

This aim of the paper is to provide an assessment of information technology and its influences on microenterprises. The literature review brings together the understanding of microenterprise through existing literature and how they are perceived in the UK economy. The survey also addresses some of the research questions highlighted within the introduction. Within such definition grounded, to then identify ways in which microenterprises establish themselves strategically and how they plan to grow is an important aspect of this research. Through the analysis of existing literature of ‘microenterprise and technology adoption’, the author would aim to draw upon how information technology affects the decisions and challenges microenterprises face when adopting IT and the capacity to help strategically grow.

To find common ground, the term ‘enterprise’ is a synonym for ‘business’, but additionally, [4] states more importantly, the term ‘enterprise’ is also about the actions of an individual that shows initiative by making something happen and taking risk through chance and investment of establishing a business, with the responsible individual commonly known as an entrepreneur [4].

Country/Area	Micro	Small	Medium	Source/Reference
Europe	1-9 employees (\$ 2m turnover)	10-49 employees (<\$10m turnover)	50-249 employees (<\$50m turnover)	[5]; [6]
UK	1-9 employees	10-49 employees	50-249 employees	[5]; [7]; [8]; [9]; [10]; [11]; [12]; [13]
Canada	1-5 employees	6-99 employees	100-499 employees	[12]; [14]
US	1-19 or 0-5	20-99 or 6-99	100-500	Small Business Administration [5]; Association for Enterprise Opportunity [12]
New Zealand	<= 5			[15]

Table 1.1 Variance of definitions of Small, Medium and Micro Enterprises

Definitions of microenterprise vary across the world, other constraints can apply but it predominately determined by size. Coincidentally many authors of publications use the definition of their locality as shown in figure 1 but also some examples like that of [16] that identified the commonly used 20 but decided to use up to 5 as a matter of choice. Therefore, there are differing opinions across researchers and organizational bodies in defining a microenterprise. The initial reasoning for categorizing business by size is way largely down to government policy making. The EU for example define SMEs this way to identify business that are able to benefit from programmes or policies specifically tailored for SMEs while on the other hand, ensures those that “exceed defined boundaries do not benefit from the support mechanisms

specifically intended for SMEs” [6]. Despite this however, it has become apparent, as identified by [7] and [5], Microenterprises are not always regarded in their own right and blended into the term “small enterprise”. Despite ‘The Department for Business, Innovation & Skills’ Sector (BIS) for the UK defines microenterprise in their 2009 report [13][31], interestingly, does not include the definition or refer to the term ‘micro’ at all in their latest report for “Business Population estimates for the UK and Regions 2012”; Instead the BIS refer to ‘small businesses’ as having between 0-49 employees [13].

Microenterprises and UK Government

Despite the existence of a well-defined term, The UK government have been highlighted by many researchers referencing them only distinguishing companies by SMEs against large companies [7], [2], [16]. By doing so, the term SME, employing fewer than 250 people refers to 99.9% of the 4.8million businesses in the UK, within that the term ‘small’ equating to 99.2% of all UK enterprises [17]. Therefore a mere 6,000 registered businesses are large with more than 250 employees accounting for only 0.1% of all UK enterprises [17]. The significant point that must be addressed, of the 4.8 million businesses registered in the UK, 4.6 million are microenterprises representing 96% of all the enterprises in the UK. Yet, so much research and emphasis is placed upon larger businesses leaving them largely ignored, blended in the shadows of the much larger definition of SMEs [5].

Inversely, [8] highlights that 95% of government funding for business skills and support goes to the 5% of UK Businesses; consisting of the large, medium and small enterprises despite the clear need for microenterprises to have more support, guidance and funding [8]. Somehow the UK Government, fail to draw upon policies and statistics in England to incorporate microenterprises. There has been a recent e-petition (December 2012) to draw signatures asking the UK government to “tell us what you are doing for Micro Enterprises not for SMEs” [18]. Our findings also analyzed that to the extent of describing the public sector e-business advisory services in the UK as poor and even potentially dangerous [10]. Given that microenterprises make a significant contribution to the economy and communities and collectively can be regarded as the “powerhouse of the economy” [23]. Studies states that “the issue is not so much that large corporations are over appreciated but microenterprises are underestimated and underappreciated” [5]. It so happens that the UK unemployment of 3.3m in 2009 would actually be 2.2m if the UK government recognized self-employed, home-office entrepreneurs which fall under microbusiness [8]. So with lack of common use of definitions and ambiguity of a small enterprise, the situation is exacerbated as Enterprises without any employees are usually ignored completely [11]; [12]. Research identified this statement in the USA stating that as much as 75% of businesses completely ignored as single-person enterprises are not even included in statistical surveys [5]. This is because as the US census bureau identifies that “most have lower overheads with no one on payroll resulting in accounting for a mere 3% of tax receipts” [5] [19]. So despite their sheer

number, [5] states that simply because of statistical prominence, they are invisible to the government as the resulting lack of access to business-supporting policies and resources.

Comparisons of Microenterprises and Small Enterprises

Research suggests that “in every case” those microenterprises have limited capital, technological and human resources comparatively to their larger counterparts [11]; [12]. Smaller firms by definition have limited internal resources and capabilities [11]; [12] while there may be some exceptions to the case, particularly in the surge of app developers and dominant companies like Plentyoffish.com (which is still today run and sustained by one person and is one of the leading global dating websites) the majority do apply to these circumstances [5] [16]. The Federation of Small Business released a report in February 2012, The FSB ‘Voice of Small Business’ member survey that helps to identify what characteristics differentiate microenterprises from small enterprises [33]. This extensive report is led by the FSB, the UK’s leading business organization representing the 200,000 owners of small businesses. A useful analytical tool to help establish a true representation of the population from the sample is the use of confidence intervals, in this report a confidence of 95% is assumed and where statistically significant differences exists, details have been included.

The following table 1.2. Illustrates the comparisons as a useful tool to grasp some of the barriers identified by the FSB. However it is very limited as it does not suggest any other factors which differs from other sources and highlighted early in the FSB report [33]. Family factors for example that are a considerable thought documented in [16], for example. The factors given to give an indication that even with little difference in definition from 1-9 employees as a micro and 10-49 as a small firm there are considerable difference. To highlight of course is staff employed but more strikingly, the net change from the previous 12 months showing that microenterprises clearly want to remain small with a flexible work force with small comparatively engaging in growing their business by as many as 20 people in full or part time. The negative change in full time staff for microenterprises may be linked to the current economic climate feeling the pinch financially but showcasing the flexibility of the work force.

		businesses	
Staff employed (average)	Total: 3.2 Full Time: 2.0 Part-time 0.8	Total: 24.7 Full Time: 14.4 Part-time: 5.1	Considerable difference
Net change in staffing levels compared to 12months before	Full Time: -3 Part Time: +2 Casual/Temp: 0	Full Time: +11 Part-time: +10 Casual/Temp: +5/+7	Considerable difference
Past Innovation: New or improved products or services in the last 2 years	66%	78%	Small consider more likely to have innovated in the last 2 years
Future Innovation: New or improved products or services in the next 2 years	60%	72%	Small more likely to innovate in the future than micro
Business Objectives – Next 12 Month	Grow: 57% Remain about the same: 29% Downsize/Consolidate business: 4%	Grow: 64% Remain about the same: 23% Downsize/Consolidate business: 5%	Micro not as inclined to grow and more likely to remain the same

Table 1.2 Factor Comparisons of Micro and Small Enterprises [33]

Microenterprises Growth

Studies highlight that despite many studies into what determines the survival and growth of microenterprises, there is no definite agreement in the literature about the key factors of success – suggesting that results vary across different studies [20]. Large corporations are driven to grow and exploit markets to the best of their abilities through profit maximization, but in contrast microenterprises are diverse in type of business and reason for existence, therefore not all microenterprises strive for growth [11]; [12]. Evidence suggests there are some ‘Growth-orientated’ microenterprises. The FSB report suggests 57% of UK microenterprise look to grow in the next 12 months. Similar studies also suggest the following characteristics towards having a positive relation to a microenterprise’s growth [33]. These characteristics are [21]:

- Educational level of an entrepreneur;
- Entrepreneurial intensity of the firm;
- Informal networking with customers and suppliers;
- Business partnering activities;
- Product innovation;
- Adoption of E-Business technologies;
- Managerial delegation;
- Focus for local markets;
- Age and size of the firm;

Authors argue and identify that growth in the traditional sense of wanting to increase the size of the firm is not the case for

Factor	Micro	Small	Note to consider
Premises Owned	27%	38%	Home working is prevalent in England
Source of Finance	Personal Savings/Inheritance, family and friends loans (£33,600)	Retained profits and bank lending (£120,000)	Significant differences in borrowing and access to the bank
Sources of business support	Seek less generally (family/friends and Informal networks more than small)	Seek more generally (predominantly higher with banks, suppliers, Solicitor, Gov’t funded support and bigger	Micro source more informal channels comparatively to small firms seeking formal advice (usually at a price) – source as a customer source shows parity

the majority of microenterprise owners [22]. The common misconception and largely Thus, increasing the size of the company by increasing the number of employees' results in a change of responsibility and tasks, detracting from the reason they started their business. Many researchers have suggested that microenterprises value the success of their business not by objective factors (profit, revenue etc.) but by subjective factors (customer satisfaction, personal motivation) [16] [24]. Researchers [15] and [20] makes an assumption and completely disregard motivation or entrepreneurial orientation to the growth of the business yet studies found and suggested it is precisely these subjective factors that drive the strategic direction and growth of a microenterprise [25] [26].

Information Technology and Microenterprises

The rapid development in Technology has fortunately resulted in a diverse range of IT to become a much more viable and cost effective solution to many business needs. With fast processing power, ability to store data easily in large quantities and have access to simple analysis tools [27]. Research studies suggest that there is little research even as wide as SMEs go in terms of understanding exactly what factors directors are influenced by in making decisions about the use of technology, let alone how it can aid the strategic direction and growth. From the unique characteristics of a microenterprise distinguishing itself from its larger counterparts, the assessment of technologies should focus on that of a microenterprise and not from a small business standpoint [15].

Microenterprises do not have the need or intensity to compete and adopt the latest technologies as many SMEs do as they are not affected as prolifically in competition [11]; [12]. Technology however has enabled reach to extend to a global scale which can and has allowed microenterprises to compete over larger corporations as IT has eliminated geographical barriers to an extent, competing on a personalized and customized level as microenterprises can still benefit from their adaptability and flexibility [29]. This research states it's limitations as it needs to identify the relationships between as mentioned before; entrepreneurial motivation and capabilities and the technology use and learning with value creation and growth in microenterprises.

Research studies has suggested that finance is the one of the critical issues that micro enterprise face when trying to adopt new IT infrastructure and solutions – the lack of recognition, funding and support by the government exacerbates the situation Invalid source specified. Research states that while there is greater emphasizes on the use of IT in the workplace, there are key differences in the use of IT comparatively between microenterprise and those that are predominantly larger [16]:

- Generally have few resources available to devote to IT
- Little control over forces that are external to the organization
- Generally do not have their own separate IT department
- Less formalized planning and control procedures for the adoption and use of IT.

Therefore it can be gathered from existing literature that small business entrepreneurs are often placed in a 'catch22' situation: knowing that IT can support their business in some way vs. lack of the expertise, resources and finance to know how it can be effectively applied. While research has shown that productivity improvements can be achieved through "learning-by-doing" in order to become more efficient and effective. In recent years the development literature suggests that when ICTs are applied in innovative ways the efficiency gains can be magnified. If applied correctly and appropriate to the social context of which such firm is within. Research identified the benefits as [30]:

- *New markets*
- *Administrative efficiencies*
- *Access to information and expertise*
- *Labor productivity can be achieved*

However, they have been recognized as difficult to achieve in microenterprises [22]. The reasons as microenterprises are already established and are high proportion in numbers suggests that there is less evidence about their growth patterns. However, what can be identified are the challenges that microenterprises face in being able to adopt and use information technology applications. Research studies describes as a "growing" business therefore SMEs looking to strategically grow. Authors are able to classify the diverse range of challenges faced by a microenterprise into the following categories of Capabilities, Resources, Access, Attitude, Context and Operations [22]. To overcome such challenges, Learning factors were the next step in research that has been undertaken as a response to this. In line with the informality of microenterprises, many studies on small businesses highlight the importance of informal learning. Hendry et al (1991) identifies 5 ways learning by: solving problems by oneself, together with colleagues, asking for help of a colleague, direct employee participation and learning new things under the responsibility of another experienced worker. It appears there is great focus of learning within the organization, this is largely due to the expense of external sources of help [11]; [12] [5]. 45% self –learn for example according to Duxbury et al. (2002) with employees with microbusiness less likely to have received employer-funded formal training; linking with the informality of the micro-workplace.

The literature review analyzed suggested that microenterprises dominate the nations' economies by number, in the UK equating for 96% of the UK's enterprises. Although some differences in defining businesses by size, microenterprises predominantly are regarded to have employees between one and nine and the fluctuations do not affect the overall consensus. Characteristically microenterprises differ greatly even from their larger counterparts of a small enterprise, particularly in how the business perceives growth. Unlike small, medium and large enterprises, the majority of microenterprises, particularly of those in the UK do not aim to grow by assessing objective factors like size and revenues but of subjective factors; to grow in terms of achieving personal value of achievement and motivation. Importantly, despite the clear differing characteristics of a microenterprise to their

larger counterparts, the government's lack of recognition of microenterprises exacerbates the problem of lacking resources, capacity to achieve their perceived value of success as funds are not fairly distributed across all SMEs.

Current literature identified that not all microenterprises are driven to succeed by becoming larger but there is a balance between the two. This is largely dependent on the business' owner's values and preferences but there are valid contrasting views of microenterprise strategic growth. The review also suggests the need to know more about the perceptions of IT. Technology has enabled microenterprises to compete against larger businesses on a personalized and customized level as they can benefit from their adaptability and flexibility but there is little research even as wide as SMEs go in terms of understanding exactly what factors directors are influenced by in making decisions about the use of technology, let alone how it can aid the strategic direction and growth.

Growth and success can and should be to an extent defined qualitatively, based on, among other things, a small enterprise owner's vision, values, lifestyle goals, and the quality of his or her relationships and contributions within and beyond the walls of business [5]. For now it is ideal to focus on what can be done without the dependency for such support and recognition from the government, particularly in these economic times as they are a big part of local communities and play a huge role in the economy regardless. Much research is focused on policy making for microenterprises and seems to not have an impact on microenterprises. So if this is the case, surely the influence of successful IT use here, reducing and perhaps automating administrative tasks for example, would enhance this success and growth of the company as it enable the owner to focus on the relationships and customer satisfaction instead of tedious administrative tasks and laborious amounts of time consumed in internal business tasks?

Where the gap in research is evident, if the microbusiness improve their internal processes and administrative tasks more efficiently through the careful use of IT, then perhaps there will be more time for the microenterprise to engage better with their customers, to identify new opportunities with products and their clients and to open new opportunities within and further away from their markets, and existing clientele. From a generic standpoint, there is empirical evidence suggesting that the best way to achieve profitability is not necessarily from cutting costs, but from improving revenues through achieving higher levels of service quality and customer satisfaction [27]. Firms that can successfully achieve both cost reduction and revenue expansion can have a positive impact on both customer satisfaction and long-term financial performance, research shows that attempting to focus on both at the same time usually fails. Therefore, the idea is to identify what can be done from an internal standpoint in terms of reducing time on process, hence also viewed as a cost reduction (in terms of wages), will then enable the microenterprise to focus on working closer with current customers and adding value to their services, and working with new clients generating new

revenues instead of been held back by administrative tasks for example.

It is this definition that can be used to help develop further research to clarify the need to differentiate microenterprises from larger businesses albeit marginally larger. Therefore, in conclusion, it is evident that microenterprises need the support to achieve their personal unique goals as defined by their own combining factors to surmount to the conceptual definition of Strategic Growth highlighted in this review. It is apparent with the lack of Government support, limited resources and the many challenges and barriers it would become apparent that research supports that IT has the potential to assist microenterprises to enable to achieve their goals and their strategic growth to obtain them.

III. RESEARCH METHODOLOGY

The research intends to use both quantitative and qualitative research methods are adopted within the research. The first method identified within the research and discussed within this paper is that of an observational case study involving observing individual participants and processes within a real business environment on a typical day of a microenterprise. This technique identified real examples of influences and challenges of Information Technology within a Microenterprise and back existing findings in current literature. Company "PMM" was involved within the study, involving the authors to go into the business and witness the general running of the business, observe behaviors and how the people within the business interact with one another and the how they utilize current IT applications. The second method of Surveys including semi-structured questionnaire consisting of questions that the participant or respondent addresses is planned for future aspect of this research, and not discussed within this paper.

IV. CASE STUDY OVERVIEW

The case study conducted presents the findings of an observational study carried out within a microenterprise. The observations consisted of monitoring the general office presence, noting any discussions of importance. Particular attention was played on observing how each employee interacted with their Information Technologies present in the business and how they undertook their general business processes, especially their administrative tasks. Highlighting the issues and problems discovered from the observations and discussions, there may be some challenges that need to be addressed. PMM or Project Management Microenterprise (name changed for this study) is a Microenterprise based in South-East England offering project management services to publishers, universities and hospital trusts. From transcripts through to final publications, PMM manages all aspects of the publication of books, journals, reports from transcript to final product including the management of authors and virtual teams to bring the project to fruition. The team have no direct contact with clients as they are worldwide – everything is practically done via email.

AS-IS Study

PMM has organically grown on very informal processes but has been very successful, even some sixteen years later there is nothing wrong with the nature of the business from an external perspective and a lot of business is ready to be taken on. However, PMM has been faced with many challenges as a result of the economic downturn and additional challenges faced through 2012, therefore a strategic choice has led to considerations of a new strategy in order to harness the potential for the business to grow. With fixed pricing in publishing and a heavy reliance on technology, much has to be done to reduce costs. As a result, the opportunity to provide PMM with possible solutions through observations conducted under this thesis has been offered that could identify and overcome the problems that exist that currently prevent PMM from becoming more efficient or grow strategically. PMM currently engage in technology for external purposes such as virtual teams and e-book publications but much is to be altered internally in order to maintain competitive and costs low.

V. ANALYSIS - ISSUES AND PROBLEMS

A. Time constraints

An issue to be aware of is PMM are bound fixed costs within their contracts and therefore have a limited number of hours their Project Managers can turn a project around. As a result each employee has to manage their timesheets by every quarter of an hour. If allocated hours are surpassed the project will be working at a loss. Therefore there more efficient PMM is, the more profitable they are.

$$\text{Total Allocated hours for Project} = \frac{\text{Number of Pages in Project} \times \text{price per page}}{\text{Project Manager cost per hour}}$$

B. Informality, external commitments and company dynamics

The Director is now looking to grow again but the culture in the company has declined since reducing their time in the office. The Director took some time out a few years ago to train in consultancy and takes on a number of responsibilities such as teaching in further and higher education sectors. But having been out of the office for some time, the informality that once created a fantastic environment is beginning to takes its toll as the team have become quite defensive with the director's presence and susceptible to any change. The Director always ensure flexibility and enables as much develop as possible with her colleagues but there is a need for perhaps some more formal process and believes they cannot employ a larger team until the dynamics in the current business changes.

C. Multiple Emails

Email is the main communication tool for engaging with clients and members of the team if out of the office. The time consuming aspect however is there may be scores of authors on one project and every single author has to be emailed individually with unique attachments. This can take a considerable about of hours to complete and a very repetitive process with relatively high margin of error.

D. Paper based projects and Administrative Burdens

Although a member of the team is devising a spread sheet to work with individual projects, all documentation, milestones, timetables and schedules are printed on paper and stored in files. As a result essential information has the potential to go missing and is not an efficient way of storing information, particularly with limited office space. This way is not particularly useful if the employees are working away from the office, posing a security risk with confidential information and is an unnecessary expense on ink and paper effecting the profitability of PMM. The information required may not be where it is thought to be (on another desk for example) and is a laborious task to trawl through paperwork to find what the individual is looking for. This method does not bear well with the potential to increase business as the office will become hard to manage with an increase of projects.

E. Internal Communications

All communication within the team is largely verbal with no means of tracking or monitoring progress or issues, it is up to the director or employees to make note of anything to do, relying on the organization and memory of each individual

F. Performance recognition

With no formal processes in place it was observed and recorded by employees there is a feeling there is little direction thus evoking a lack of motivation with issues going unsolved such as alleviating tensions between employees. As a result potential is not met and efficiencies dwindle. There are no targets or deliverables set with no method of recording performance. In addition there is no Structured Appraisal and no rewards for project completion or achievement related targets. Employees spoke needing "guidelines, objectives and targets".

G. Data entry

Each member of staff have their own Spread-sheet to record their projects based on a master copy. However there are a number of issues identified. There is no consistency amongst staff, each with their own style and data entry is not consistent across projects. There is therefore no means of keeping track of all projects relying on the user to keep track of deadlines, this issue is exacerbated as it common to maintain 10-12 projects at any one time.

H. Internal Server

PMM have an internal server they rely heavily as this stores all of their files and data required for projects. This allows for file sharing between colleagues on their computer as well as offering an FTP facility to upload files from their virtual teams. However, the reliability is not excellent and while they have a backup it only updates once a day, so any work undertaken during a day's work and a power cut occurs, it is likely much of the data is lost. Security is also an issues especially with so many external members involved with projects from publishers, typesetters and authors.

VI. KEY STRATEGIC CHALLENGES

As a microbusiness the problems and issues PMM face are difficult to overcome. The challenges identified are the following:

A. Resources

- Lack of Money - PMM lack finance to invest their own resources into the business and currently only generates enough to cover costs;
- Lack of Time - Projects consume a full working day; Family is priority so takes precedence alongside work; External commitments consumes work hours and therefore is made up another time; The Director nor employees have minimal time during office hours and outside the office environment available to seek ways to enhance their business
- Lack of Support - Support usually only comes at a price exceeding the added value of return
- Lack of Information - Not easier to find simple solutions to a unique company without investment

B. Capabilities

- Lack IT knowledge - The director and employees are proficient users of IT and benefit from the latest desktop PCs however they lack know-how of how to optimize existing processes with the use of Information technology
- Inadequate IT user skills - although one member of staff is attempting to enhance their current spreadsheet, the user is only going by basic experience of spreadsheet software and with the hours put in, overall improvements will not greatly enhance the business.
- Inability to resolve technical issues - the current in-house server requires a third-party call-out team if problems exist due to lack of technical know-how;
- Understanding the benefits of IT - use of IT is critical in the business yet paper based process still consume much the time, space and expense without seeking the potential of the IT that is available;
- Limited Planning Ability - PMM have limited time to plan a strategy as they are constantly responding to the demands of their clients and have limited time to do so
- Not knowing what your IT actually does - this challenge is not so much an issue as they do use all technology available to them
- Not knowing what their current IT potential is - with the latest Desktop PCs, fast internet and use of desktop publishing software, there is potential to exploit if it is known;

C. Attitude

- Lack of engagement with colleagues - despite less than 5 people work in the company office there is lack of engagement of some colleagues as they merely do the job in hand and do not seek to improve the company;
- Resistance to technology - use of paper may suggest there may be slight resistance to utilising their

technology - evident in data input styles in spreadsheets;

- Lack of value and personal incentives - no performance recognition or appraisal
- Lack of awareness - staff are not aware of some cost effective solutions to enhance the business (cloud-based data storage)
- Lack of trust - it would appear with the reduction of hours the director is in the office there is an increase in self-management and less engagement and tensions between colleagues - resulting in lack of trust
- Lack of confidence - initial reactions were apprehensive of finding ways to improve the business before solutions were identified
- Cultural Factors - as a globally communicating microenterprise, language over email may be misinterpreted, national holidays may not be accounted for;

D. Access

- Inadequate hardware and software - In-house server may not be adequate in data size, speed and reliability - especially if looking to grow the business; new desktop publishing software is no inadequacy to PMM
- Poor IT infrastructure - network is sound but the in-house server is a reliability and security issue that should be addressed as it is proving an expensive maintenance issue;

From the case study findings it was identified that PMM currently to not optimize the business to the best of their ability and will not be able to expand successfully until current processes and visions are standardized and the use of Information Technology is utilized effectively

VII. PROPOSED SOLUTION

With PMMs current choice of strategy of retrenchment in the short-term leading to growth in the medium-long term, arisen from the challenges faced through 2012. Suggestions addressed is with the inclusion of some solutions addressed as a result of the observations/ consultation the aim short-term is to reduce costs by utilizing IT more effectively. In the short term, this will increase productive capacity. An added value to PMM of using a more consolidated approach to IT will be the provision of a training and induction programme containing standard and accepted practice which will be explicit, allowing specific SMART objectives. Therefore a conceptual model has been constructed that could address the outcome of the consultancy report. As the time spent on processes reduces as a result of IT solution development, it is hoped that it will have an exponential opportunity to increase business. The next stage of the research is to further investigate and propose a conceptual model which will aim to address the challenges identified within the research and that

of the use of Information Technology within Microenterprises allowing them to attain strategic growth.

VIII. CONCLUSION

The following study aimed at the identification and analysis of few important research questions. The key research question addressed within this study was to identify the need of Microenterprises having greater awareness in society differentiating from Small-Medium Enterprises (SMEs). Despite little recognition to differentiate microenterprises generally in society, with an established definition and considerable differing characteristics of their slightly larger counterparts 'Small enterprises', more should be done to promote microenterprises as an individual entity, albeit within an SME but for the sake of the Government identifying microenterprises that suffer greatly from financial challenges. Microenterprises want to added value and increase business if there was a greater influence of IT on their strategic growth which would in turn boost the economy. The research also aims to address to evaluate if all microenterprises are driven to become bigger businesses or they prefer to remain 'Micro' for financial and strategic reasons. Our literature findings and that of case study analysis suggests that diverse ranges in microenterprise objectives and therefore a near fifty-fifty split look to become bigger whereas the remaining prefer to stay as microenterprise. A greater majority of microenterprises however look to remain micro overall disputing the generalization 'all businesses grow big'.

Our study also looked on evaluating the current perceptions of IT and its impact on microenterprises aiming to pursue their current and future growth strategy within the context. Results suggested that there is a positive perception of Information technology holding a critical success factor of their strategies with 78% of surveyed participants, without it many businesses would cease to exist and many identify a realm of benefits Information Technology has to offer. From literature the list of barriers, number of barriers had been identified that hold greater importance, largely the management of number of factors including: financial, time, data and people management [22]. Our findings also suggested that there is great potential with the right balance of suitable solutions, time and skills adopted; the solutions in the next chapter will provide more evidence of this.

REFERENCES

- [1] Johnson, G., Scholes, K. & Whittington, R., 2008. *Exploring Corporate Strategy*. 8th ed. Harlow: Pearson Education Limited.
- [2] Tyler, R., 2010. *Explosion of micro companies recorded*. Available at: <http://www.telegraph.co.uk/finance/yourbusiness/8004994/Explosion-of-micro-companies-recorded.html>
- [3] Experian, 2012. *British businesses battle back from the brink of bankruptcy*. Available at: <http://press.experian.com/United-Kingdom/Press-Release/british-businesses-battle-back-from-the-brink-of-bankruptcy.aspx>
- [4] Riley, J., 2012. *Starting a business - what is meant by "enterprise"?* http://www.tutor2u.net/business/gcse/enterprise_what_is_enterprise.html
- [5] Walters, J.S., 2002. *Big Vision, Small Business: 4 Keys to success without growing big*. 7th ed. San Francisco: Berrett-Koehler Publishers Inc.
- [6] European Union, 2003. L 124. *Official Journal of the European Union*, 46, p.36.
- [7] Devins, D., Johnson, S., Gold, J. & Holden, R., c2001. *Management Development and Learning in Micro Businesses: a 'missing link' in research and policy*. Report prepared for: Research and Evaluation Unit Small Business Service. Leeds: Leeds Metropolitan University Policy Research Institute.
- [8] SFEDI, 2009. *Six Killer Facts: in micro enterprise learning and support*. [PDF] Small Firms Enterprise Development Initiative Ltd Available at: http://www.sfedi.co.uk/sfedi-news/six_killer_facts.pdf.
- [9] Welsh Government, 2012. *Micro-Business Task and Finish Group Report: Jan 2012*. Annual Report. London: Welsh Government Micro-Business Task and Finish Group.
- [10] Simpson, M. & Docherty, A.J., 2004. E-Commerce adoption support and advice for UK SME's. *Journal of Small Business and Enterprise Development*, 11(3), pp.315-28.
- [11] Davis, C.H., Lin, C. & Vladica, F., 2006. Internet Technologies and e-business solutions among small and medium-sized enterprises (SMEs) in Atlantic Canada - Patterns of Use, Business Impacts, and demand for Support Services. In *Proceedings of the 7th World Congress on Management for e-Business*. Halifax, 2006.
- [12] Davis, C.H. & Vladica, F., 2006. *The value of Internet Technologies and e-business solutions to microenterprises in Atlantic Canada*. PhD Thesis. New Brunswick, Canada: University of New Brunswick Faculty of Communication and Design, Ryerson University.
- [13] BIS, 2011. *BIS Small business survey 2010*. [Report] Department for Business Innovation & Skills; https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/32228/11-p74-bis-small-business-survey-2010.pdf
- [14] Government of Canada, 2012. *Industry Canada - SME Research and Statistics*; From: <http://www.ic.gc.ca/eic/site/061.nsf/eng/02714.html>
- [15] Jouvenaux, M., 2007. Micro-enterprises, technology and e-commerce in New Zealand. In *11th Pacific-Asia Conference on Information Systems*. Auckland, 2007. PACIS.
- [16] Burgess, S., 2002. *Challenges and Solutions*. London: Idea Group Publishing
- [17] BIS, 2012. *Business Population Estimates for the UK and Regions 2012*. [Statistical Release] Department for Business, Innovation & Skills Available at: <http://www.bis.gov.uk/assets/BISCore/statistics/docs/B/12-92-bpe-2012-stats-release.pdf>
- [18] HM Government, 2012. *e-petition "tell us what you're doing for micro enterprises nt for SMEs"*. [Online] HM Government Available at: <http://epetitions.direct.gov.uk/petitions/18396>.
- [19] United States Census Bureau, 2012. *Statistics about business size (including small business) from the U.S. Census Bureau*. [Online] Available at: <http://www.census.gov/econ/smallbus.html>.
- [20] Lasch, F., Le Roy, F. & Yami, S., 2007. Critical growth factors of ICT start-ups. *Management Decision*, 45(1), pp.62-75.
- [21] Berry, A. & Perren, L., 2001. The role of non-executive directors in UK SMEs. *Journal of Small Business and Enterprise Development*, 8(2), pp.215-32.
- [22] Wolcott, P., Kamal, M. & Qureshi, S., 2008. Meeting the challenges of ICT adoption by micro-enterprises. *Journal of Enterprise Information Management*, 21(6), pp.616-32.
- [23] Rashid, M.A. & Al-Qirim, N.A., 2001. E-Commerce Technology Adoption Framework by New Zealand Small to MEdium Size Enterprises. *Res. Lett. Inf. Math. Sci*, 2, pp.63-70.
- [24] Chau, S. & Pedersen, S., 2000. Small is beautiful: the emergence of new micro business utilising electronic e-commerce. In *Australian Conference on Information Systems*. Brisbane, Australia, 2000.
- [25] Bellu, R.R., 1993. Task Role motivation and attributional style as predictors of entrepreneurial performance: femal sample findings". *Entrepreneurial and Regional Development*, 5, pp.331-44.
- [26] Chell, E., Haworth, J.M. & Brearly, S.A., 1991. *The Entrepreneurial Personality: Concepts, Cases and Categories*. London: Routledge.
- [27] Rust, R.T. & Espinoza, F., 2006. How Technology advances influence business research and marketing strategy. *Journal of Business Research*, 59, pp.1072-78.
- [28] Rust, R.T., Zeithaml, V.A. & Lemon, K.N., 2004. Customer-centered brand management. *Harvard Business Review*, 82(9), pp.110-8.
- [29] Chew, H.E., Levy, M. & Ilavarasan, V., 2011. The Limited Impact of ICTs on Microenterprise Growth: A Study of Businesses owned by

- Women in Urban India. *Information Technologies & International Development*, 7(4), pp.1-16.
- [30] Qureshi, S., 2005. How does information technology effect development? Integrating Theory and Practice into a Process Model. In *Proceedings of the eleventh Americas conference on Information Systems*. Omaha, NE, 2005.
- [31] BIS, 2011. *BIS Small business survey 2010*. [Report] Department for Business Innovation & Skills https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/32228/11-p74-bis-small-business-survey-2010.pdf
- [32] BIS, 2012. *Business Population Estimates for the UK and Regions 2012*. Department for Business, Innovation & Skills <http://www.bis.gov.uk/assets/BISCore/statistics/docs/B/12-92-bpe-2012-stats-release.pdf>
- [33] FSB, 2012. *The FSB 'Voice of Small Business' Member Survey*. Birmingham: Federation of Small Business.

Binarization and recognition of characters from historical degraded documents

Bency Jacob

Department of Computer Engineering
Sinhgad Institute of Technology
Lonavla, India
bencyjac@gmail.com

Mr. S.B. Waykar

Department of Computer Engineering
Sinhgad Institute of Technology
Lonavla, India
sbwaykar@gmail.com

Abstract— Degradations in historical document images appear due to aging of the documents. It is very difficult to understand and retrieve text from badly degraded documents as there is variation between the document foreground and background. Thresholding of such document images either result in broken characters or detection of false texts. Numerous algorithms exist that can separate text and background efficiently in the textual regions of the document; but portions of background are mistaken as text in areas that hardly contain any text. This paper presents a way to overcome these problems by a robust binarization technique that recovers the text from a severely degraded document images and thereby increases the accuracy of optical character recognition systems. The proposed document recovery algorithm efficiently removes degradations from document images. Here we are using the ostus method ,local thresholding and global thresholding and after the binarization training and recognizing the characters in the degraded documents.

Keywords—binarization, denoising, global thresholding, local thresholding, thresholding

I. INTRODUCTION

The old handwritten documents, manuscripts, printed books which included various historical books, old papers which were written by our ancestors. All these documents are of very importance for us today. But due to the less proper management these documents are no longer in readable form. Now a days these historical documents are presented in digital form for various purposes[1]. Digital era has made paper documentation an extinct process for everything today is done with the help of computers.

The goal of this project is the secure data Academic libraries, institutions and historical museums pile-up or preserve documents in storage areas. Our work in this paper contributes to documents safe and efficient preservation in its original state throughout the years and their unconditional exploitation to researchers, a major issue for historical documents collections that are poorly preserved and are prone to

degradation processes, see Documents digitalization, allows access to wider public, while cultural institutions and heritage organizations create local or national digital libraries accessed through the internet. Our work concentrates on basic techniques used for image enhancement and restoration, denoising and binarization. The entire system is implemented in visual environment. The document is recovered and characters are recognised.

To analyze the document, its image is binarized before processing it. It is nothing but segmenting the document background & the foreground text. For the confirmation of document image processing task an accurate document image binarization technique is a must. After years of studies in document image binarization, the thresholding of degraded document images is still found to be a challenging task because of the high inter/intra variation between the text stroke and the document background across various document images. The stroke width, stroke brightness, stroke connection, and document background vary in the handwritten text within the degraded documents. Moreover, bleed through degradation is observed in historical documents by variety of imaging outputs. For most of the existing techniques many kinds of document degradations, it is still an unsolved problem of degraded document image binarization due to the document thresholding error. A document image binarization technique presented in this paper is an extended version of an existing local maximum minimum method [5].

II. RELATED WORK

To convert the image into its binary format, many thresholding techniques are used in practices. As many degraded documents, lot of variation in image pattern, global thresholding cannot be a better approach for the degraded document binarization and thus adaptive thresholding is better approach.

Generally Otsu's thresholding is used as global thresholding. But cannot be used for image having large number of variation, but if windowing is used and then for each window Ostus is applied then the thresholding can be enhanced and

new array can be evaluated for further processing to get a clear segmentation of text from the background.

[3] Other approaches have also been reported, including background subtraction texture analysis, recursive method decomposition method, contour completion Markov Random Field, matched wavelet, cross section sequence graph analysis, self-learning, Laplacian energy user assistance and combination of binarization techniques But these method are often complex for analyses.

OCR mainly deals with improving the efficiency and accuracy. All efforts in OCR technology concentrate on this property. The OCR technology for Roman scripts or Indian, follow the same basic methodology of pre-processing, segmentation, feature detection and extraction, and classification as referred by [8]. In this approach vertical and horizontal projections are used for line, word and character segmentation which obtain a performance result of 93%. The [9] proposes an OCR system including pre-processing by binarization and size normalization by trial and error methods. They perform segmentation using projection profile and report a recognition rate of 87%. Much of the exploration on OCR system's efficiency improvement, in Indian context, has revolved around the exploration of using Otsu's method for preprocessing. The Otsu's thresholding algorithm is the basic thresholding technique used popularly for binarization in most works.

The [10] discusses Otsu thresholding method gives a better binarization result in degraded documents Global thresholding algorithms are usually faster as they use a single threshold based on the global histogram of the gray-value pixels of the image. This method is an improvement to existing methods to create a novel and effective way to binarize historical documents.

The [11] proposes a method for binarization of image using its texture features. They use Otsu thresholding iteratively to produce a threshold values, and texture feature associated with each threshold are retrieved using run length algorithm. They report an improvement of 8.1% over the original Otsu's algorithm. A new method for the binarization of the image other than Otsu's is proposed by [12] in which binarization is done using Morphological operators. Morphological operators are very efficient in that it performs image binarization effectively by taking care of complement color combination patches in the image.

The Morphological operator serves better in case of noisy corpus, especially when border noise is present in abundant. Statistics show that thresholding method fails completely in case of fore mentioned cases. This is done by using two filters based on two operations dilation and erosion [12]. It gives a much better result than Otsu's binarization method by removing of the background noises much effectively.

III. PROPOSED METHOD

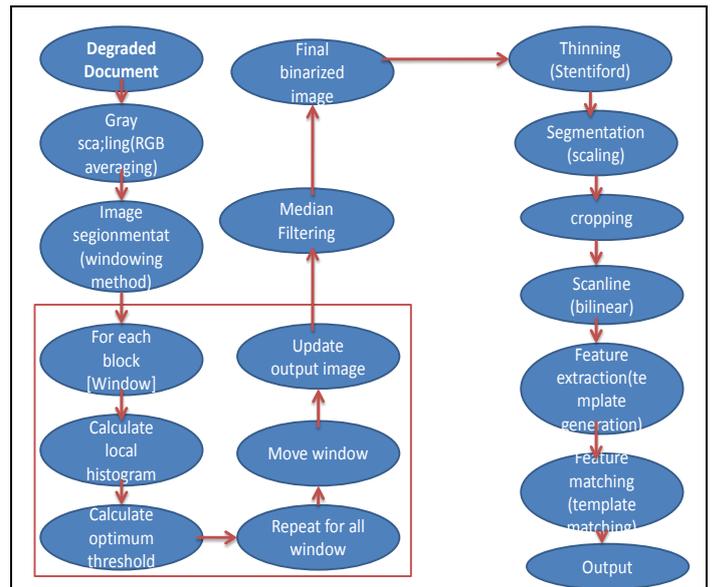


Fig1. System Flow

There are two important parts of system are :

1. Binarization
2. OCR

Here we are first converting the image to grayscale, then dividing the image into windows and on each window we calculate the local histogram, calculate the optimum threshold this is done on each window and then using the media filter the noise is cleared and the binarized image so obtained is thinned, segmented, cropped, scaled, then feature extracted and using the training and recognising the characters in the text are recognized.

For a given a degraded document image, R, G, B colours are separated from a given coloured image and then each color is ANDed with 0xff to obtained the 8 bit binary value of each colour (R, G, B).

Then After separating each colour gray scaling which is 8 bit binary value is obtained. Thresholding is applied on the gray scale image. In this paper basically window based thresholding is applied and then Otsu's is applied over which window to obtain.

Threshold value for each window. This is done in the preprocessing stage. Edges are then detected using sobel algorithm to the threshold image. And finally filtering is done using median filter. Proposed method is simple and requires minimum parameter tuning. To enhance the quality performance of the technique OCR is added at the last.

Formula for Grey Scaling is $(GS) = (R+G+B)/3$.

Thresholding will be applied to gray scale image value i.e. only two values will be generated either Black Or white I.e. gray scale value > threshold then the pixel will turned to white

& If gray scale value < threshold then the pixel will be turned black

A. Windowing technique

After getting the gray scale image window function is applied to the grayscale image.

For image blur Window can

be of size 3 by 3, 5 by 5 or 9 by 9. Less will be the window size, less blur; vice versa. Windows width and height (means how many pixels in X and how many pixels in Y)

Eg- $(100 \times 100) \times (\text{window size})$ $100 \times 100 \times (3 \times 3)$ W * H * (size of the window)

While traversing through each window Otsu's will be applied to each window so that threshold is obtained for each window. This is done because the image has large variation.

Otsu's alone cannot be used for the complete image as it gives the single global threshold value, so if the image has variation information will be lost.

B. Detecting the edges

To detect the edges of the image Sobel algorithm is used. It has standard matrix for horizontal and vertical edges. They can be named Hx and Hy, where Hx is used to find horizontal edges and Hy is used to find vertical edges of the text.

C. Median Filter

After creating foreground pixel map, some morphological post processing operations such as erosion, dilation and closing are performed to reduce the effects of noise and enhance the detected regions. Noise is also called salt and pepper noise and is removed by using median filter.

While going through the text in the window if neighboring pixel does not have any overlap edge of the text then it will be treated as noise and converted to white (i.e. 1). So after applying the filter we will get the text in readable form.

Then the image is then thinned using Stentiford algorithm. Then segmented to get each character and then cropped accordingly. Later the scaling technique is used to scale the individual character and then the characters are trained and features are extracted and features are matched and the characters are recognized.

IV. MATHEMATICAL MODEL

A. Otsu's

In Otsu's method exhaustive search for the threshold that minimizes the intra-class variance (the variance within the class).[]

Defined as a weighted sum of the two classes.

$$\sigma_w^2(t) = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t) \quad \dots 1$$

Weight ω_i are the probabilities of the two classes separated

by threshold t and t variances of these classes. Otsu shows that minimizing the intra-class variance is the same as maximizing inter-class variance.

$$\sigma_b^2(t) = \sigma^2 - \sigma_w^2(t) = \omega_1(t)\omega_2(t) [\mu_1(t) - \mu_2(t)]^2$$

Which is the expression in terms of class probabilities ω_i and class means. The class probability is computed from the histogram as t

$$\omega_1(t) = \sum_0^t p(i)$$

While the class mean μ_i

$$\mu_1(t) = \left[\sum_0^t p(i)x(i) \right] / \omega_1$$

Where $x(i)$ is the value at the center of histogram bin. Similarly you can compute $\omega_2(t)$ and μ_2 on the right hand side of the histogram for bins greater than t . The class probabilities and class means can be computed iteratively. This idea yields an effective algorithm.

In computer vision and image processing, Otsu's method is used to automatically perform clustering-based image thresholding,[1] or, the reduction of a graylevel image to a binary image. The algorithm assumes that the image contains two classes of pixels following bi-modal histogram (foreground pixels and background pixels), it then calculates the optimum threshold separating the two classes so that their combined spread (intra-class variance) is minimal.[2]

Otsu's Algorithm

- 1 Compute histogram and probabilities of each intensity level
- 2 set up initial $\omega_i(0)$ and $\mu_i(0)$
3. step through all possible threshold $t=1..$ maximum intensity

- a. update ω_i and μ_i
- b. compute $\sigma_b^2(t)$

- 4. Desired threshold corresponds to the maximum $\sigma_b^2(t)$
- 5. you can compute two maxima (and two corresponding threshold) $\sigma_{b1}^2(t)$ is the greater max and $\sigma_{b2}^2(t)$ is the greater or equal maximum
- 6. desired threshold = (threshold1 + threshold2) / 2

$a_1 \leq a_2 \leq a_3 \leq \dots \leq a_n \in R$ the new value of intensity of a pixel (i, j) of an image I is given by:

$$median(A) = \begin{cases} a_{\frac{n+1}{2}}, & \text{if } n \text{ is odd} \\ \frac{1}{2} \left(a_{\frac{n}{2}} + a_{\frac{n}{2}+1} \right), & \text{if } n \text{ is even} \end{cases} \dots 1$$

B. STENTIFORD ALGORITHM

Stentiford Thinning Algorithm [7]

It uses a set of four 3 x 3 templates to scan the image. Shows these four templates.

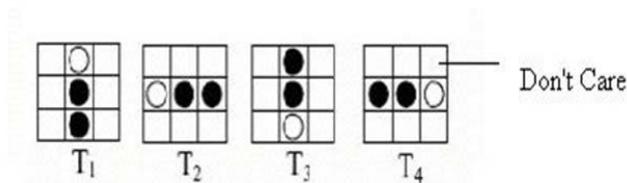


Fig2. Stentiford Thinning Algorithm

The Stentiford Algorithm can be stated as following:

1. Find a pixel location (i, j) where the pixels in the image match those in template T1. With this template all pixels along the top of the image are removed moving from left to right and from top to bottom.
2. If the central pixel is not an endpoint, and has connectivity number = 1, then mark this pixel for deletion. Endpoint pixel: A pixel is considered an endpoint if it is connected to just one other pixel. That is, if a black pixel has only one black neighbour out of the eight possible neighbours.
3. Repeat steps 1 and 2 for all pixel locations matching T1.
4. Repeat steps 1-3 for the rest of the templates: T2, T3, and T4. T2 will match pixels on the left side of the object, moving from bottom to top and from left to right. T3 will select pixels along the bottom of the image and move from right to left and from bottom to top. T4 locates pixels on the right side of the object, moving from top to bottom and right to left.
5. Set to white the pixels marked for deletion.

Median Filter [3]

Median filter is a non linear filter.

For

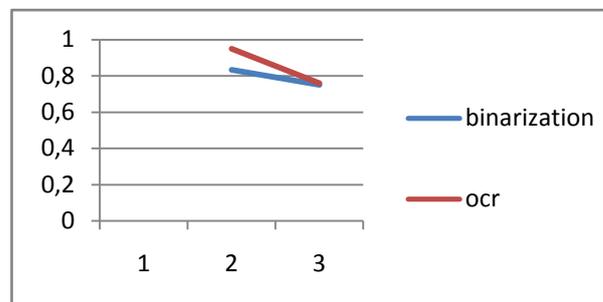
$A \{a_1, a_2, a_3, \dots, a_n\}$, and

V. RESULTS AND DISCUSSION

To make the text in the degraded document understandable and to train and recognize the characters from the degraded document. The precision and recall values are as shown in the results below. Which shows results of how many characters from the degraded documents are properly binarized and how many characters are recognized.

DataSet Name	Actual Objects	Retrieved Objects	Correct Retrieved Objects
binarization	20	18	15
ocr	25	20	19

Confusion Matrix		Class 1	Class 2
	Class 1	15	1
	Class 2	13	19
	Precision		
binarization	0.833333333		0.75
ocr	0.95		0.76
Total	1666667	0	0.755
Accuracy Percentage	0.755		



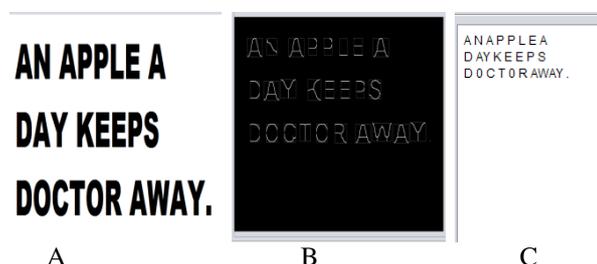
Testing outcome



A.



The outcome of the proposed method which is the binarization result A showing the input image, B the output of the base paper, C the outcome of the proposed method.



Outcome of the character recognition method A showing the input image, B the trained image and C showing the recognized characters.

V. CONCLUSION

In this paper, is proposed a robust approach for removing degradations and recovering text printed and handwritten scanned document images by binarization and training and recognizing the characters which is a challenging task. Most of the document analysis and recognition works reported are on good-quality documents. But still it remains a highly challenging task to implement a character reconision that works under all possible conditions and gives highly accurate results. Elaborate studies on poor-quality documents are not much undertaken by the scientists in the development of script independent OCR. Experiments should be made to observe the effect of poor quality paper as well as noise of various types, and take corrective measures.

Acknowledgment

I submit my gratitude and sincere thanks to my guide Prof.S.B.Waykar, Head of Computer Department S.D. Baber, who has been very concerned and has aided for all the material essential for the dissertation work and preparation of this thesis report, He helped me to explore this vast topic in an organized manner and provided me with all the ideas on how to work towards a research oriented venture.I am thankful to our ME Co-ordinator Prof. M. S. Chaudhari, for his unwavering moral support and motivation during the entire course of the project. I would also like to thank our Principal

Dr. M. S. Gaikwad who encouraged us and created a healthy environment for all of us to learn in best possible way. I would like to thank all the staff members of our college and technicians for their help.

References

- [1]. Bolan Su, Shijian Lu, and Chew Lim Tan, Senior Member, IEEE 'Robust Document Image Binarization Technique for Degraded Document Images', APRIL 2013
- [2] NTOGAS, NIKOLAOS and VENTZAS, DIMITRIOS, 'A BINARIZATION ALGORITHM FOR HISTORICAL MANUSCRIPTS,' in Proc. 28th Int. Conf. VLDB, Hong Kong, China, 2002,pp. 155-16612th WSEAS International Conference on COMMUNICATIONS, Heraklion, Greece, July 23-25, 2008
- [3] Brij Mohan Singh and Mridula, 'Efficient binarization technique for severely degraded document images,' CSIT DOI 10.1007/s40012-014-0045-5
- [4] Arie Shaus, Eli Turkel and Eli Piasezky, 'Binarization of First Temple Period Inscriptions - a Performance of Existing Algorithms and a New Registration Based Scheme,' 2012 International Conference on Frontiers in Handwriting Recognition
- [5] Maya R. Gupta and Nathaniel P. Jacobson, Eric K. Garcia, 'OCR binarization and image pre-processing for searching historical documents,' Elsevier Received 28 October 2005; received in revised form 27 February 2006; accepted 28 April 2006
- [6] Sayali Shukla, Ashwini Sonawane, Vrushali Topale, Pooja Tiwari, 'Improving Degraded Document Images Using Binarization Technique(ISSN : 2277-1581)Volume No.1', INTERNATIONAL JOURNAL OF SCIENTIFIC TECHNOLOGY RESEARCH VOLUME 3, ISSUE 5, May 2014.
- [7] wikipedia, www.google.com'
- [8] Bansal V, Sinha RMK, "A complete OCR for printed Hindi text in Devanagari script", Sixth International Conference on Document Analysis and Recognition, 2011, pp. 800-804.
- [9] Desai A, Malik L, Welekar R, "A New Methodology for Devanagari Character Recognition", International Journal of IT, Vol. 1, 2011, pp. 626-632.
- [10] Liu Y, Srihari SN, "A recursive Otsu thresholding method for scanned document binarization", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, 1997, pp. 540-544
- [11] Brodic D, Milivojevic DR, Tasic V, "Preprocessing of Binary Document Images By Morphological Operators", In MIPRO 2011 Proceedings of the 34th International Convention, 2011, pp. 883-887.
- [12] Shafait F, Breuel TM, "A simple and effective approach for border noise removal from document images", IEEE 13th International Multitopic Conference INMIC, 2009, pp. 1-5

Adaptive analysis of characteristic nodes using prediction method in DTN

First A. Yoon-hyung Dho, Second B. Kang-whan Lee

Abstract— In this paper, we propose an algorithm that analysis characteristic nodes to select efficient relay nodes using information of network environment. The weight factor in proposed algorithm is changeable as following network environment as node density. Consequently, we select adaptive relay node to rough environment networks. Existing Delay Tolerant Networks(DTNs) routing algorithms have problems the large latency and overhead because of deficiency of network information in unsteady network. We have to solve this problem, predict future network using node state information and apply weight factor that changes according to the network environment. Thus, selected relay nodes by proposed algorithm work efficiently in unstable and stressed network environment. Simulation results show that enhanced performance as overhead, delivery ration, average latency compared to existing DTN routing algorithms.

Keywords— DTN, Ad-hoc, Location-based, Prediction-based.

I. INTRODUCTION

Delay Tolerant Network(DTN) is a network architecture that is designed to communicate to unstable end-to-end network[1]. Existing wireless network as WLAN set the pre-routed path using heterogeneous network, and send message. However, interruption of communication is occurred frequently in environment that lost infrastructure, applying to conventional protocol is difficult as TCP/IP. DTN uses the store-and-forward message delivery based approach to solve this problem, end-to-end connection is to preserve the message via the relay node, even in unstable conditions enable communication.

Due to the characteristic of DTN which insufficiency of whole network view, routing path can't be pre-selected. Consequently, performance of network is changed depending on whether the selected relay nodes. Existing prediction-based DTN routing method determines the routing path, and selects the relay nodes according information of node's location history or local connectivity. These existing DTN routing methods select relay node in insufficiency states of network environments less reliable due to the frequent change in performance.

In this paper, we propose method that analysis state

Yoon-hyung Dho is with the Korea University of Technology and Education (email: zephyrus@koreatech.ac.kr)

Kang-whan Lee is with the Korea University of Technology and Education.(email: kwlee@koreatech.ac.kr)

information of node's characteristic and select efficiency relay node using dynamic weight factor of selecting relay nodes according to network environment to solve these problem.

II. RELATE WORK

DTN routing protocols are significantly defined by the deterministic and stochastic routing protocols[2]. Deterministic routing protocols assume situations that previously known location and mobility information of future nodes. As a result, the method of transmitting messages can be applied by using the information of each node such as the mobility information. Typically deterministic protocols is Oracle-based method[3] that transmit message using oracle information and [4] that predict path of nodes using space-time graph. In this paper, we assume situation that flexible and estimated network. Thus deterministic routing protocols are different to our purpose.

Stochastic routing protocols assume that unknown network environment, message should be considered to be passed when and what. Stochastic routing protocols include Epidemic[5], Spray and Wait[6], Prophet[7], Bubble rap[8] and so on.

Epidemic routing and many other stochastic routing method copy and forward message all neighbor nodes those does not have same message. This way, such as by a message transmitted has been known to be the most effective for the delivery of messages in uncertain network environments, each node in the DTN a copy of the message that this is limited because it is has a limited buffer. Consequently, performance of message transmit ratio is lowered and overhead is occurred highly on account of generated network overload.

These routing method copy and forward message to all neighbor nodes without determination for efficient diffusion of message. Therefore problems of message transmission are occurred for deficiency of node's buffer and usable relay nodes. In addition, if node density is highly enough in network number of copied message is increasing rapidly. Thus problems occur in the use of network resource.

Prophet and Bubble rap set efficiency relay nodes using nodes history information to solve these problem. But if in sparse network, accurate analysis is difficult using node history in these method due to lower contact opportunities between the nodes. Since recent research has been actively to solve these problem of stochastic routing protocols[9].

This paper is a part of these study, we propose a method that predict current and future state of network using analysis of node's history and dynamic weight factor of relay node

selection to inducing selection of the efficiency relay nodes.

III. PROPOSED METHOD

Proposed method in this paper, basically adapted flooding based message forward method as stochastic routing method similar to Epidemic and Spray and Wait. Existing flooding based routing method assume that mobility of nodes is random. But the actual node's mobility has predetermined information of destination and directionality accounting on finality of nodes.

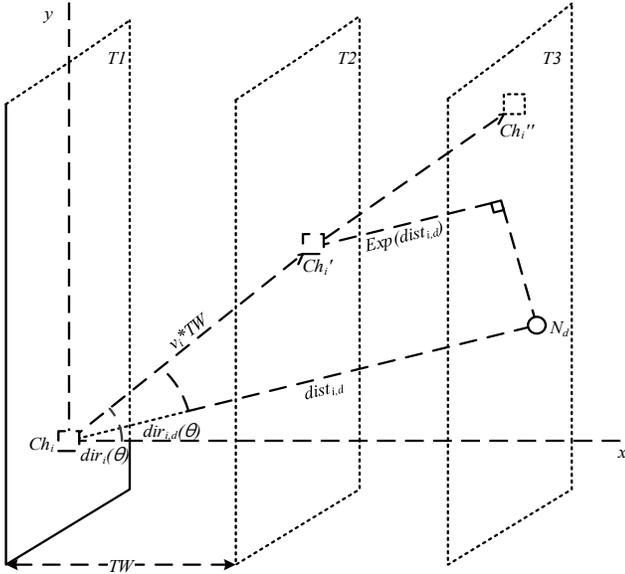


Fig. 1 Illustration of locational information for prediction of node's state

Node works as follow when contact neighbor nodes. First, Replace the attribute information table in the node type of each node with messages. Next, compares the attribute information of each node selects the node with the possibility to communicate with the relay node and the destination. And compares the message with the selected relay node. Last, copies the message to the selected relay node are not the same.

The probability P_i which neighbor node i is selected as relay node is computed by (1).

$$p_i = (w_{dir} \cdot \cos(dir_{i,d}\theta) + (w_v \cdot \frac{v_i}{MAX(v)})) + (w_{dist} \cdot \frac{Rt_d}{dist_{i,d}})$$

$$dir_{i,d}\theta < \frac{\pi}{2}$$

$$dist_{i,d} < Rt_d$$

$$(1)$$

$$w_{dir} + w_v + w_{dist} = 1$$

$$MAX(p_i) = 1$$

In (1), w_{dir} is weight factor of direction, w_v is weight factor of velocity, w_{dist} is weight factor of distance. And $dist_{i,d}$ is segment between node i and node d . The $dir_{i,d}(\theta)$ is degree of angle

within segment $dist_{i,d}$ and v_i . The Rt_d is transmit range of destination node. TW(Time Window) is constant value for the purpose of estimation of node's movement. Weight factors in (1) are changed dynamically according to network environment for selection of more efficient relay nodes.

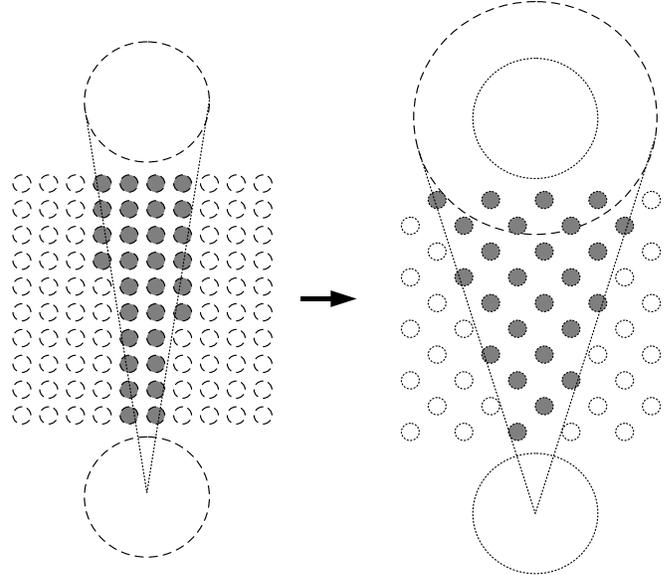


Fig. 2 Illustration of relation between relay node selection and node density

A. Direction weight factor

W_{dir} is effected by node density in network. Through $w_{dir} \cdot \cos(dir_{i,d}(\theta))$ in (1), we know that decide degree of angle to select relay node. As a result, relation of W_{dir} and relay node selection is represented as figure 2. Figure 2 show that if the greater node density, W_{dir} is decreased and if node density is small, W_{dir} is increased. Through the change of W_{dir} , we decrease overhead rate according to node density and keep the efficient degree of relay node selection.

But in DTN, the overall node density is unknown since characteristic of DTN as unstable end-to-end connection. So node density for correcting W_{dir} is determined movement distance of each node and number of neighbor node for period of time. Figure 3 show relation between node movement range and node density. Through figure 3, we know that node density is resolvable according to number of neighbor node and transmit range in TW . Resolved node density is represented by (2).

In (2), The Base is distance of node movement in TW . The Area is dimension that constituting the distance of node movement and transmit range. The $N_{neighbor}$ is number of neighbor nodes. Selectable number of relay node by degree of angle are increased according to resolved node density $density(t)$ of network though (2). t_0 is time of last TW . As a result, adapted w_{dir} for efficient selection of relay node is represented as (3).

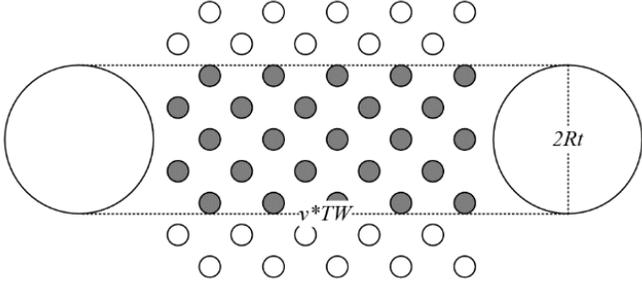


Fig. 3 Illustration of relation between node movement range and node density in network.

$$\begin{aligned}
 Base &= v \cdot TW = \int_{t_0}^t v(t) dt \\
 Area &= \frac{\pi}{2} (Rt_{t_0}^2 + Rt_t^2) + \int_{t_0}^t Rt(t) dt \cdot base \\
 &= \frac{\pi}{2} (Rt_{t_0}^2 + Rt_t^2) + \int_{t_0}^t Rt(t) dt \cdot \int_{t_0}^t v(t) dt \\
 density(t) &= \frac{\sum_{t=t_0}^t N_{neighbor}}{Area} \\
 &= \frac{\sum_{t=t_0}^t N_{neighbor}}{\frac{\pi}{2} (Rt_{t_0}^2 + Rt_t^2) + \int_{t_0}^t Rt(t) dt \cdot \int_{t_0}^t v(t) dt}
 \end{aligned} \quad (2)$$

$$w_{dir}(t) = w_{dir}(t_0) \cdot (1 - density(t)) \quad (3)$$

The adapted W_{dir} though (3) is decreased or increased by $Density(s)$. Consequently, number of selected relay node are decreased when node density of network is increased.

B. Velocity weight factor

W_v is effected by distribution of message. Distribution of message is decided by TTL(Time To Live) of message. If distribution of message is greater, the message is estimated to old message and remaining TTL will be shorter. Figure 4 show distribution of message according to time in flooding based communication method. As a result, messages those have short TTL are represented as low distribution but have long TTL are represented as high distribution.

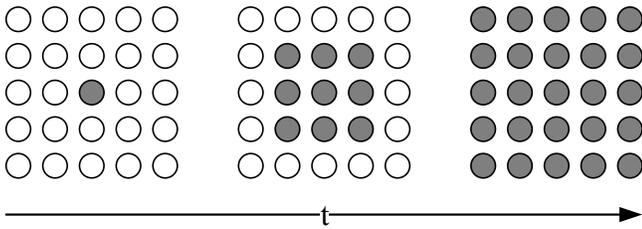


Fig. 4 Illustration of message distribution according to time for flooding based message forward method.

So messages that has high distribution are forwarded by relay nodes with high velocity since high probability of which have

short TTL. Therefore, nodes are processed to decide message distribution

- Old message count (M_{old}): if nodes are in contact with the same message, M_{old} is increased 1 in these message.
- Fresh message count (M_{fresh}): if nodes are in contact and copy new message each other, M_{fresh} is increased 1.

Through using M_{old} and M_{fresh} , each node decide message distribution MD. MD is represented by (4).

$$M_D(t) = \sum_{t=t_0}^t \frac{M_{old}(t)}{M_{old}(t) + M_{fresh}(t)} \quad (4)$$

According to (4), if old message count is high, M_D is increased. If old message count is low, M_D is decreased. Consequently, adapted W_v is represented by (5).

$$w_v(t) = w_v(t_0) \cdot M_D(t) \quad (5)$$

Though (5), W_v is increased when network message distribution is high. As a result, probability of selecting relay nodes which have high velocity is increased. Therefore the nodes have message which have short TTL select fast relay nodes easier and message delivery ratio is increased.

C. Adapted Probability of relay node selection

As a result, adapted probability of relay node selection P_i is represented to (6) using (3) and (5). Through (6), P_i is decided by dynamic weight factor W_{dir} and W_v that is changeable according to resolved network environment. Therefore, each node select context and efficient relay nodes using P_i . Consequently, network overhead and overhead rate is decreased and the message delivery ration is maintained.

$$\begin{aligned}
 p_i(t) &= (w_{dir}(t_0) \cdot (1 - density(t)) \cdot \cos(dir_{i,d}\theta) \\
 &\quad + (w_v(t_0) \cdot M_D(t) \cdot \frac{v_i}{MAX(v)}) \\
 &\quad + (w_{dist} \cdot \frac{Rt_d}{dist_{i,d}}) \\
 dir_{i,d}\theta &< \frac{\pi}{2} \\
 dist_{i,d} &< Rt_d
 \end{aligned} \quad (5)$$

$$\begin{aligned}
 w_{dir} + w_v + w_{dist} &= 1 \\
 MAX(p_i) &= 1
 \end{aligned}$$

IV. SIMULATION AND RESULTS

In this section, we compare to Epidemic routing that is existing stochastic routing protocol, prediction based routing[10] that is not using the proposed method and prediction based routing using proposed method using simulation. Simulation environment is shown table 1.

TABLE I. SIMULATION ENVIRONMENTS

Network size	900x600(m ²)
Simulated time	12(h)
Time To Live	2(h)
Number of nodes	100,200,300
Transmit range of nodes	5m
Transmit range of base station	50m
Velocity of nodes	1~5m/s
Mobility model	Random Way Point Model

A. Overhead

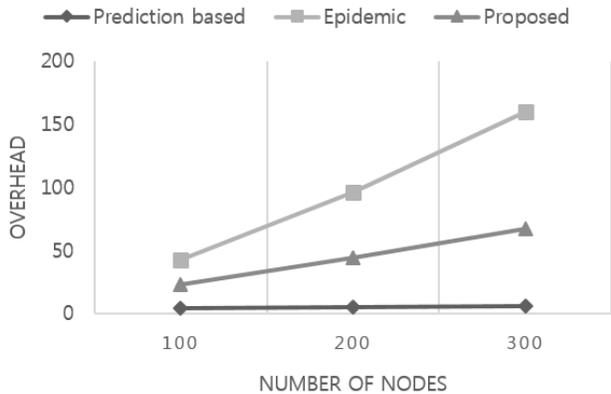


Fig. 5 Simulation results of overhead as change of node's number.

Figure 5 is graph that compare overhead to existing routing protocols and routing protocol using proposed method according to increasing number of nodes. Figure 5 show overhead of routing protocol using proposed method more 38% compared to existing prediction based routing protocol. But compared to Epidemic routing protocol, overhead is decreased to average 47% according to node increases. And overhead rate is reduced to 50% in accordance with the node increases. As a result, figure 5 is shown that proposed method is more effective in dense network environment.

B. Figures and Tables

Large figures and tables may span both columns. Place figure captions below the figures; place table titles above the tables. If your figure has two parts, include the labels “(a)” and “(b)” as part of the artwork. Please verify that the figures and tables you mention in the text actually exist. **Please do not include captions as part of the figures. Do not put captions in “text boxes” linked to the figures. Do not put borders around the outside of your figures.** Use the abbreviation “Fig.” even at the beginning of a sentence. Do not abbreviate “Table.” Tables are numbered with Roman numerals.

C. Delivery ratio

Figure 6 is graph that compare message delivery ration to existing routing protocol and routing protocol using proposed method. Delivery ratio of prediction based routing protocol using proposed method show high delivery ratio in sparse

network environment unlike existing prediction based routing

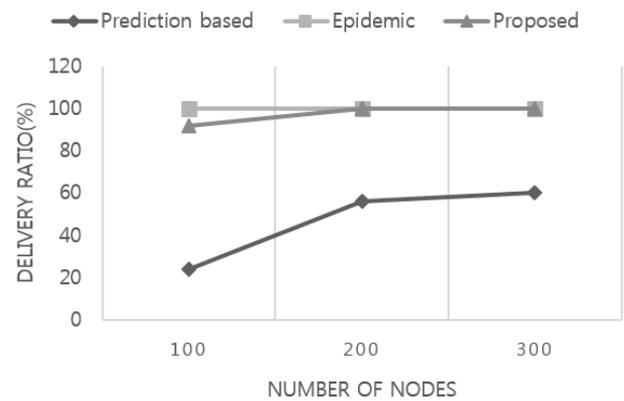


Fig. 6 Simulation results of delivery ratio as change of node's number.

protocol. And delivery ratio is increased similar to Epidemic routing protocol according to nodes increases. Consequently, delivery ration of prediction based routing protocol using proposed method is increased from twice to four times compare to existing prediction based routing protocol.

D. Average latency

Figure 7 is a graph comparing the average delay time according to the number of nodes increases. Average latency of prediction based routing protocol using proposed method is short in sparse network environment unlike existing prediction based routing protocol. And proposed method shows the

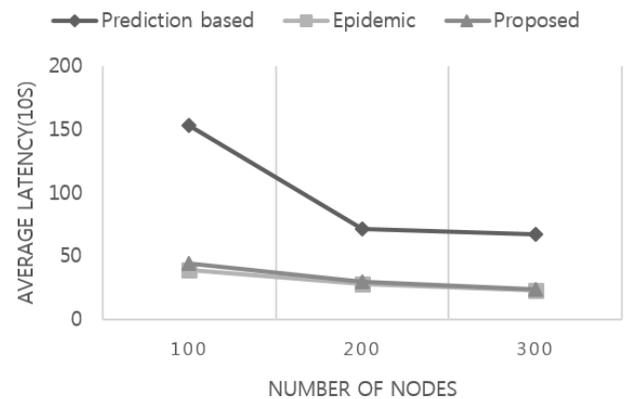


Fig. 7 Simulation results of average latency as change of node's number.

average latency decreased from 50% up to 80% compared to the existing prediction based routing protocol.

Prediction based routing protocols using the proposed method is shown high delivery ration and short average latency compared to existing prediction based routing protocol in sparse network environment. Furthermore, overhead is reduced 50% compared to Epidemic routing protocol. And overhead ratio is reduced according to nodes increases. Consequently, proposed

method complement the disadvantage of existing prediction based routing protocol in sparse network environment using dynamic weight factor according to network environment. And disadvantage of Epidemic routing protocol as high overhead is complemented without performance penalty of delivery ratio and average latency.

V. CONCLUSION

DTN is purposed to solve problem of unstable network environment as space, deep sea. But existing DTN routing focus message delivery ratio and neglect efficiency network communication. Consequently, existing DTN routing method is too hard to apply to actual condition of network environment.

In this paper, we propose method that select more efficiency relay nodes in DTN to solve these problem. For select efficiency relay nodes in DTN, proposed method analysis node state information in node's history and appreciate network environment. So proposed method adapt initiatively to unstable network environment using dynamic weight factor of relay node selection. Furthermore, node's mobility is methodical differently to Random Way Point Model[11] in simulation. Therefore, the closer to real network environment, accuracy of prediction of node's movement and actual performance will be better. If you continue to further study considering relay node selection according to the buffer and the energy of the node, it will be more reliable configuration of the routing in DTN.

ACKNOWLEDGMENT

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning(2014H1C1A1066391), Korea, under the Specialized Co-operation between industry and partially supported by the Education and Research Promotion Program of KUT.

REFERENCES

- [1] Delay Tolerant Networking research group <http://www.dtnrg.org>
- [2] Zhensheng Zhang, Routing in Intermittently Connected Mobile Ad Hoc Networks and Delay Tolerant Network: Overview and Challenges, IEEE Communications Survey and Tutorial, 2006, pp.24-37.
- [3] Thrasylvoulos Spyropoulos, Konstantinos Psounis, Cauligi S. Raghavendra, Single-copy routing in intermittently connected mobile networks, In Proc. Of IEEE Secon, 2004.
- [4] Huai-En Lian, Chien Chen, Je-Wei Chang, Chien-Chung Shen, Rong-Hong Jan, Shortest Path Routing with Reliability Requirement in Delay Tolerant Networks, Future Information Networks, 2009, First International Conference on.
- [5] Amin Vahdat, David Becker, Epidemic Routing for Partially-connected Ad Hoc Networks, Duke University, 2000, Technical Report CS-2000-06.
- [6] Thrasylvoulos Spyropoulos, Konstantinos Psounis, Cauligi S. Raghavendra, Spray and Wait : An Efficient Routing Scheme for Intermittently Connected Mobile Networks, ACM Workshop on Delay Tolerant Networking, 2005, pp.252-259.
- [7] Anders Lindgren, Avri Doria, Olov Schele'n , Probabilistic Routing in Intermittently Connected Networks, ACM SIGMOBILE Mobile Computing and Communications Review, 2003, Vol.7, no.3, pp.19-20.

- [8] Pan Hui, Jon Crowcroft, Eiko Yoneki, Bubble rap: social-based forwarding in delay tolerant networks, Proc. MoniHoc, 2008, pp.241-250.
- [9] Pei-Chun Chen, Kevin C. Lee, Mario Gerla, Jérôme Härrı, GeoDTN+Nav : Geographic DTN routing with navigator prediction for urban vehicular environments, Mobile Network, 2010, vol.15, no.1, pp.61-82.
- [10] Yue Cao, Zhili Sun, Naveed Ahmad, Haitham Cruiskshank, A Mobility Vector Based Routing Algorithm for Delay Tolerant Networks Using History Geographic Information, IEEE Wireless Communications and Networking Conference: mobile and Wireless Networks, 2012.
- [11] Tracy Camp, Jeff Boleng, Vanessa Davies, A survey of mobility models for ad hoc network research, Wireless Communications & mobile Computing, 2002, Special Issue on Mobile Ad Hoc Networking: Research, Trends and Applications, vol.2, no.5, pp.483-502.

Cyber diversity for security of digital substations under uncertainties: assurance and assessment

E. Brezhniev, V. Kharchenko, J. Vain, A. Boyarchuk

Abstract— Cyber diversity is proposed as one of the general principles of fault- and intrusion-tolerant digital (smart) substations connected to Nuclear Power Plant (NPP). The approach for substation diversity assessment and assure to decrease risks of Common Cause Failures (CCFs) and Cyber CCF is suggested in this paper. The new metrics of smart substations diversity assessment are introduced allowing estimating the diversity required to decrease risks of CCFs including cybersecurity risks. Smart substation's attributes of diversity (design, signal, functional, equipment and others) are also considered in the paper.

Keywords - cyber diversity, NPP, smart substation, common cause failures

I. INTRODUCTION

A. Motivation

Power electricity industries are being widely transformed, driven by the need for more energy, scarcity of natural resources and global warming. One of the most promising technologies is the smart grid technology which combines a traditional electric power grid (PG) with an "intelligent" information and communications technologies to produce a smarter power system.

Nuclear power occupies a unique position in the debate over global climate change as the only carbon-free energy source. Nowadays it is already contributing to world energy supplies on a large scale, and has potential to be expanded if the challenges of safety, nonproliferation, waste management, and economic competitiveness are addressed, and technologically fully mature. So it might be concluded that NPP is an intrinsic part of future smart grid.

The substations provide links between NPP and PG. They are extremely strategic to NPP safety. Compared to other systems in an electric utility network, the smart substation has the highest density of valuable data needed to operate and manage a smart grid. Unreliable substation equipment and insufficient cyber security introduce new risks to NPP safety. A successful attack on one of these substations could have fatal and expensive consequences.

Smart substations main assets are not only physical facilities, but also information, databases and software applications, different intelligent electronic devices (IEDs) such as breaker controllers, voltage regulators, remote terminal units (RTUs), programmable logic controllers (PLCs). These IEDs are important cyber assets of digital substation.

Substation state of operability could be compromised by IEDs common cause failures (CCFs) which could occur at any substation level, cause the substation unavailability and as a result introduce new risks to NPP safety. Hardware CCFs are failures (or unavailable states) of substation equipment due to a shared cause. NPP instrument and control (I&C) failures analysis proves that CCFs are

significant contributors to I&C incidents. For example, 450 failures (out of 3000) fall on multiple I&C failures during 564 reactor-years [1].

According to [2], Industrial Control Systems (ICS) have the common cyber vulnerabilities. These vulnerabilities are divided on three general categories such as: the vulnerabilities inherent in the ICS product, vulnerabilities caused during the installation, configuration, and maintenance of the ICS and the lack of adequate protection because of poor network design or configuration. For example, through bad coding practices and improper input validation, access can be granted to attackers allowing them to have unintended functionality or privilege escalation on the systems. Examples of improper input validation identified are within buffer overflows, boundary checking, and code injection. It means that besides hardware or software CCFs (stipulated by application the similar software versions) digital substations' IEDs might be prone to cyber common cause failures (CCCFs). Cyber CCFs might be determined as events when cyber assets' availability, confidentiality and integrity are compromised within a specified (short) time interval. The reasons are the common cyber vulnerabilities, tough coupling within networks between IEDs which might lead to security violation due to human errors, shared input data equipment, environmental events (flooding, storm) and cyber attacks, in particular, attacks on joint component vulnerabilities. Thus substations with critical loads, such as NPPs, should be given the highest level of importance in respect of their cyber security considering CCCFs as well. The higher level of cyber security of smart substation with critical load might be achieved through implementation of substations' diversity when IEDs with similar functionalities are different and less vulnerable to the same shared cause.

B. Related works analysis

The paper [3] has performed the common vulnerabilities assessments on a large variety of systems, and for each assessment, it tailors the assessment and methodology to provide the most value to the customer. All vulnerability identification activities are focused on enabling the identification and remediation of the highest risk ICS cybersecurity vulnerabilities rather than the collection of data for statistical purposes. It also gives the recommendations for ICS vendors and owners on how to reduce the common vulnerabilities of ICS systems including creation of a security culture, enhancing ICS test suites, applying patches, redesigning network protocols for security, etc. It might be unfeasible to implement the joint plan for common vulnerabilities reduction considering such business issues as competition among vendors, lack of coordination, etc.

In paper [4] the authors consider the application of “defense in depth” (DD) for cyber security assurance of electrical distribution systems. DD is a strategy of integrating technology, people, and operations capabilities to establish variable barriers across multiple layers of an organization. These barriers include electronic countermeasures such as firewalls, intrusion detection software/components, and antivirus software, coupled with physical protection policies and training. The cost of implementation of such strategy is not considered in this paper.

In paper [5] the common vulnerabilities of control systems are also considered. The paper describes the generalized trends in vulnerabilities observed from the assessments, as well as typical reasons for these security issues and the introduction to an effective mitigation strategy. Many of these vulnerabilities result from deficient or nonexistent security governance and administration, as well as budgetary pressure and employee attrition in system automation. It is also mentioned that DD concept should be used to cyber security assurance of cyber components.

In [6] diversity approach to NPP I&C safety assurance and gap-based security assessment technique and tools is described and analyzed.

It might be noted that the application of smart substation variety (diversity) is not considered as a means to decrease CCFs of IEDs including cyber CCFs. There is also no analysis of impact of substation diversity on probability of event when these substations with critical load are failed due to one shared cause. This event is determined as substations CCFs when all of them got unavailable within one short time interval. Substation cyber security assessment is characterized by high level of uncertainties existing due to high volume of qualitative data presented as experts’

opinions. To improve the validity of any cyber security estimates of smart substation the quantitative and qualitative data shall be considered and analyzed.

C. Goal of the paper

The main objective of this paper is to present the approach for cyber diversity assessment of substation under uncertainties and evaluate how this diversity can decrease the risks of CCCFs induced by common vulnerabilities among IEDs. To consider the impact of substation diversity on risks of substations’ CCCFs the application of Bayesian Belief Network (BBN) is proposed in this paper.

II. THE STAGES OF DIVERSITY ASSESSMENT OF SMART SUBSTATION WITH A CRITICAL LOAD

Diversity is one of the general principles used to decrease hardware vulnerability against CCF and provide dependability of I&C [6]. Diversity is the general approach used for decreasing CCF risks of NPP I&C systems. Differences in equipment, development and verification technologies, implemented functions, etc. can mitigate the potential for common faults [7].

Typically, there are three substations that provide the power supply for NPP. It is presumed in the future that NPP will be connected to PG through the same amount of smart substations. It is important to consider all possible risks which might occur within this interaction, analyze them and mitigate as well. Joint decision making between electrical utilities and NPP shall be also considered.

There are many risks factors for smart substation. The list of these risk factors includes the following: human failures (on different substation levels), hardware failures, software failures, cyber security issues, external events (see Fig.1).

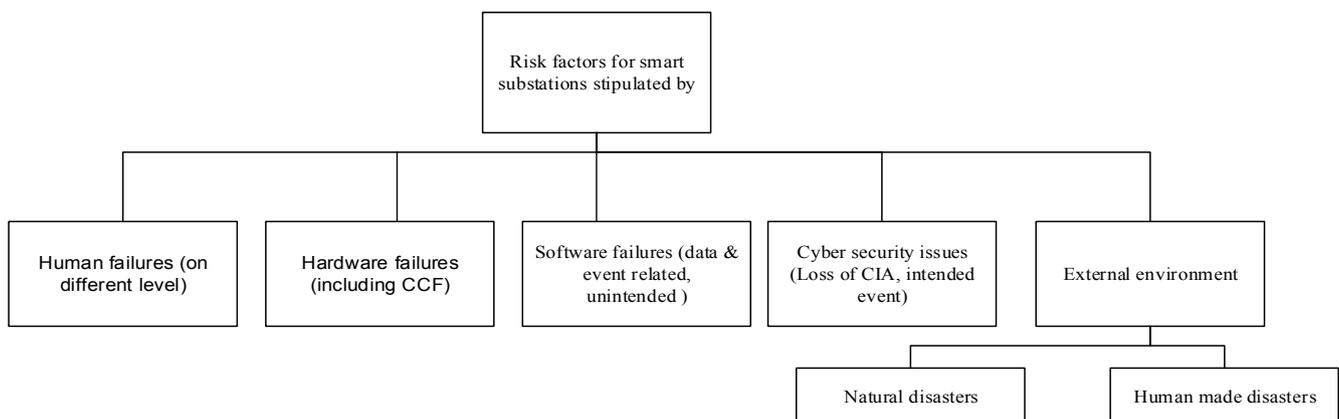


Fig 1. The risks factors for smart grid substation

Due to high level of coupling and interconnections between IEDs they might be prone to hardware and software failures.

Hardware CCFs are subset of dependent failures in which two or more IEDs fault states exist at the same time, or within a short time interval, as a result of a shared cause (root-cause). The short time interval is determined by time of system response on failure. IED software failure is considered as an inability of its program to continue signal acquisition, consolidation, processing due to erroneous logic

(usually systematic failures). There are many techniques to provide the software reliability. Among them are the software fault-tolerance techniques such as software redundancy. Software redundancy is achieved by incorporating some additional software components that are not exactly identical but they are similar in functionality. *Common cause software failures* are subset of software failures in which two or more IEDs fault states exist at the same time, or within a short time interval, as a result of a shared cause (software failure as a root-cause).

Cyber CCF might be determined as an event when IEDs' availability, their data confidentiality and integrity are compromised within a specified (short) time interval.

When an intruder gains the access to substation IEDs then the possible consequences might include: shut down of the substation or any portion of the subsystem controlled by the compromised IEDs; change IEDs settings to degrade their reliability and, subsequently, the power supply service that provided to NPP; gather control and protection settings data that could be used in a subsequent attack on other similar substation; plant malicious code that could later trigger a delayed or coordinated attack, etc.

All of these events put the new risks to NPP connected to this smart substation. NPP and power utilities' owners should cooperate with aim to reduce all possible risks caused by loss of external (for NPP) power supply. It means they should make the joint decision on selection of smart substations with the high level of diversity. The cost issues are to be considered as well. In this case the diversity is taken as CCFs mitigation techniques. Considering the IEDs high importance it is presumed that IEDs are significant contributor to substation vulnerabilities and substation cyber security might be achieved by implementation of diverse IEDs within different substations. The approach suggested for selection of smart substations with diverse IEDs deals with qualitative aspects represented in qualitative terms by means of linguistic variables. Computing with words (CW) has been applied as a computational basis to linguistic decision making of complex situations [6].

To select the most diverse grid substations, using the diversity criteria and evaluate the similarity (difference) between IEDs, expert should take into consideration the compelling evidence. Based on these evidences experts evaluate the difference (similarity) between similar IEDs (from different substations) using the linguistic terms: SAME (S), NEARLY SAME (NS), DIFFERENT (D).

A. The formation of diversity strategies set

The following diversity strategies of smart substations implementation are considered in this paper:

- Strategy S_{11} - all smart substations and their IEDs are similar. One vendor develops and produces all substations with no difference in IEDs design, manufacturing, cyber issues, etc.;
- S_{21} - all substations are different and produced by different vendors;
- S_{31} - all substations' IEDs are produced by one vendor but there are some differences between them.

B. The diversity strategies set's expertise

During this stage experts are supposed to fill the comparison matrixes to evaluate the similarities (differences) between the IEDs in term of hardware and cyber aspects. The expert is supposed to compare the IEDs with similar functionalities from different substations and select the most different between them.

If the particular IED for the first substation is determined then it is required to compare it with the possible alternatives for IEDs with similar functionalities from second and third substation. If the substation automation controller, for example, OM600, the grid automation controller of ABB, is selected for the first substation, according to S_2 strategy, this IED is compared with C264 from Alstom Grid (IED1), GE's D25 from General Electric (IED2) and SICAM (IED3) AK from Siemens. The expert is required to assign the weight of each criterion.

The criterion's weight might be expressed either as linguistic value (Low, Medium, High) or any numerical values from [0, 1]. For sake of simplicity the weight of criterion is presented as a scalar value.

Table I represents the example of diversity assessment for the strategy S_{21} (hardware aspects).

TABLE I. CHECK LIST FOR IEDs CCF (HARDWARE VULNERABILITIES ASPECT)

Vulnerabilities criterion	W_k , weight of criterion	IEDs		
		IED1	IED2	IED3
Design				
System Layout/Configuration	0,2	NS	D	NS
Component Internal Parts	0,23	NS	D	D
Design team	0,13	D	D	D
Design procedures	0,24	NS	D	NS
V&V procedures	0,2	NS	D	NS
Manufacturing				
Manufacturing method, and material	0,13	NS	NS	NS
The manufacturing staff	0,27	D	D	D
The same quality control procedure	0,6	NS	NS	NS
Installation				
Installation method, and material	0,33	NS	NS	NS
The Installation staff	0,41	D	D	D
The quality control procedure	0,26	D	D	NS
Operation				
Operation method, and material	0,4	S	NS	S
The Operation staff	0,32	D	D	D
The quality control procedure	0,28	NS	NS	D
Maintenance				
Maintenance/ Test/Calibration Schedule	0,21	NS	D	NS
Maintenance/ Test/Calibration Procedure	0,31	NS	D	NS
Maintenance/Test/Calibration Staff	0,48	D	D	D

Table II represents the example of diversity assessment for the set of strategies S₂₁ (cyber aspect).

TABLE II. CHECK LIST FOR IEDs CCF (CYBER VULNERABILITIES ASPECT)

Vulnerabilities criterion	W _k , weight of criterion	IEDs		
		IED1	IED2	IED3
Design				
The coding practices	0,12	NS	D	D
The security requirements	0,21	NS	D	D
The security testing procedure	0,13	NS	NS	NS
The vendor	0,15	D	D	D
The tools used	0,19	S	S	NS
The security culture	0,2	NS	S	D
Installation				
The installation procedure	0,43	D	D	D
The installation team	0,35	D	D	D
The installation tool	0,22	NS	NS	NS
Operation				
The communication links	0,51	S	NS	NS
The port security on network equipment	0,49	NS	D	NS
Configuration				
Patch management procedure	0,22	NS	NS	D
Encryption procedure	0,13	D	NS	NS
Authentication procedure	0,65	D	D	D

The expert is proposed to use linguistic values to evaluate all possible IEDs' alternatives for substations.

In this paper, we shall use labels represented by triangular fuzzy numbers. A triangular fuzzy number, denoted by M = <m, α, β>, has the membership function:

$$\mu_M(x) = \begin{cases} 0, & \text{for } x \leq m - \alpha \\ 1 - \frac{m-x}{\alpha}, & \text{for } m - \alpha < x < m \\ 1, & \text{for } x = m \\ 0, & \text{for } x \geq m + \beta. \end{cases} \quad (1)$$

The point m, with membership grade 1, is called the mean value and α, β are the left hand and right hand spread of M respectively.

For example, we assign the following semantics to the set of three terms:

$$\text{NS} = (0, 0,25, 0,5), \text{ S} = (0,25, 0,5, 0,75), \\ \text{D} = (0,5, 0,75, 1).$$

C. The aggregation stage

During this stage all linguistic values provided by experts are aggregated to obtain a collective assessment for the IED's alternatives. It is provided by calculation of the fuzzy diversity score D_{ij} as an arithmetic mean:

$$D_{ij} = \left(\frac{1}{t} \sum_{k=1}^t w_k \times m_{ij}^t, \frac{1}{t} \sum_{k=1}^t w_k \times \alpha_{ij}^t, \frac{1}{t} \sum_{k=1}^t w_k \times \beta_{ij}^t \right) \quad (2)$$

where w_k – weight of k criterion; <m_{ij}^t, α_{ij}^t, β_{ij}^t> – a triangular fuzzy number that represents one of linguistic values {S, NS, D} assigned by tth expert for S_{ij} diversity

strategy. D_{ij} represents a difference between two IEDs. The more value D_{ij}, which corresponds certain diversity strategy S_{ij}, the more diverse both IEDs.

Using the best-fit method [8], the obtained fuzzy diversity score D_{ij} for each IEDs can be mapped back to one (or all) of the defined linguistic terms (SAME, NEARLY SAME, DIFFERENT). The method uses the distance between fuzzy diversity score, represented by fuzzy triangular number for each IEDs and each of the initial linguistic terms to represent the degree to which obtained score, is confirmed to each of them. The distance between the obtained fuzzy diversity score D_{ij} and the expression SAME, NEARLY SAME, DIFFERENT is defined as follows:

$$d_{ij}^{(r)}(D_{ij}, \text{SAME}) = \left[\sum_{j=1}^3 (\mu_{D_{ij}}^j - \mu_{\text{same}}^j)^2 \right]^{\frac{1}{2}}$$

$$d_{ij}^{(r)}(D_{ij}, \text{NEARLY SAME}) = \left[\sum_{j=1}^3 (\mu_{D_{ij}}^j - \mu_{\text{NS}}^j)^2 \right]^{\frac{1}{2}} \quad (3)$$

$$d_{ij}^{(r)}(D_{ij}, \text{DIFFERENT}) = \left[\sum_{j=1}^3 (\mu_{D_{ij}}^j - \mu_{\text{different}}^j)^2 \right]^{\frac{1}{2}}$$

Hence, each IED is characterized by 3-tuple <d_{ij}^(r), d_{ij}^(r), d_{ij}^(r)>, where d_{ij}^(r) – a distance between obtained fuzzy diversity score and corresponding linguistic term (SAME, NEARLY SAME, DIFFERENT).

It should be noted that each d_{ij}^(r) (j = 1, ..., J, where J – number of possible alternatives classified as type of S_i strategy, I – number of strategy type) is an unsealed distance. The closer D_{ij}, is to the rth expression, the smaller d_{ij}^(r) is. More specifically, d_{ij}^(r) is equal to zero if D_{ij}, is just the same as the rth expression in terms of the membership

functions. In such a case, D_{ij} should not be evaluated to other expressions at all due to the exclusiveness of these expressions. To embody such features, new indices need to be defined based on $d_{ij}^{(r)}$ ($r = 1, 2, 3$).

Suppose $d_{ij}^{(3)}$ is the smallest among the obtained distances for D_{ij} , and let $\alpha_{i1}, \alpha_{i2}, \alpha_{i3}$ represent the reciprocals of the relative distances between the identified fuzzy diversity score D_{ij} , and each of the defined linguistic terms with reference to $d_{ij}^{(3)}$ (smallest distance). Then, $\alpha_{ij}^{(r)}$ ($r = 1, 2, 3$) can be defined as follow:

$$\alpha_{ij}^{(r)} = \frac{1}{d_{ij}^{(r)}}, r = 1, 2, 3. \tag{4}$$

If $d_{ij}^{(3)} = 0$ it follows that $\alpha_{ij}^{(3)}$ is equal to 1 and the others are equal to 0. Then, $\alpha_{ij}^{(r)}$ ($r = 1, 2, 3$) can be normalized by:

$$\beta_{ij}^{(r)} = \frac{\alpha_{ij}^{(r)}}{\sum_{r=1}^3 \alpha_{ij}^{(r)}}, r = 1, 2, 3 \tag{5}$$

Each $\beta_{ij}^{(r)}$ represents the extent to which D_{ij} belongs to the r th defined linguistic terms. Thus, $\beta_{ij}^{(r)}$ could be viewed as a degree of confidence that obtained fuzzy scores for all diversity strategies S_{ij} belong to the r th defined linguistic terms.

Results obtained for the selection of the most diverse substation controller to decrease the cyber vulnerability of smart substation with critical load are presented in the table III. The IED1 is the most diverse IED to OM600 in respect to cyber vulnerabilities.

TABLE III.RESULTS OBTAINED FOR ALL IEDs CONSIDERED IN EXAMPLE

IED alternatives	Degree to which D_{ij} belongs to the initial terms		
	S	NS	D
IED1	0,12	0,39	0,49
IED2	0,36	0,28	0,38
IED3	0,33	0,63	0,04

C. The exploitation stage

During this stage all IEDs' alternatives are ranked by using the collective linguistic assessment obtained in the previous stage, taking into account the cost of each IED, C_{ij} . The rational diverse strategy could be found with the following criterion:

$$S_{ij} = \operatorname{argmax} \frac{\beta_{ij}^{(r)}}{C_{ij}}, \tag{6}$$

where $\beta_{ij}^{(r)}$ represents the extent to which D_{ij} belongs to the r th defined linguistic terms; C_{ij} - cost of S_{ij} reduced to $\sum C_{ij}, i, j$ - indexes of alternatives.

The main aim of all stages described above is decrease the IDEs cyber common vulnerabilities of substations with

critical load. The more diversity in particular IEDs the more diversity is achieved among this type of cyber security assets. All smart substation cyber assets should be evaluated in the same way to provide the highest level of cyber security. The result of this approach is the set of diverse smart substation with critical load.

D. BBN as a basis for Cyber Common Cause Failure assessment of substations

The connections between NPP and smart substations are represented as BBN with nodes (NPP reactor, safety systems and substations) and edges as the power lines. BBNs are very effective for modeling situations where some information is already known and incoming data is uncertain or partially unavailable. BBN that represents links between reactor unit, its safety systems (RPS, RCIC) and on site, off site power supply is given in Fig. 2. Construction and assessment of BBN parameters was performed using Netica 5.12 tool. The Netica APIs are a family of powerful Bayesian Network toolkits.

BBN allows evaluating CCFs of substations that provide the power supply to NPP. When successful cyber attack on smart substations IDEs is launched then if there is no any substation diversity (all substations IEDs are the same) the probability of CCFs of all substations is 0,729. All substations are prone to the same threats and have the common cyber vulnerabilities. This scenario describes the situation when the attacker (terroristic organization) performs the successful attempt to compromise the IEDs cyber security and as a result to make the substation to be not operable. It seems to be realistic while all substations are similar.

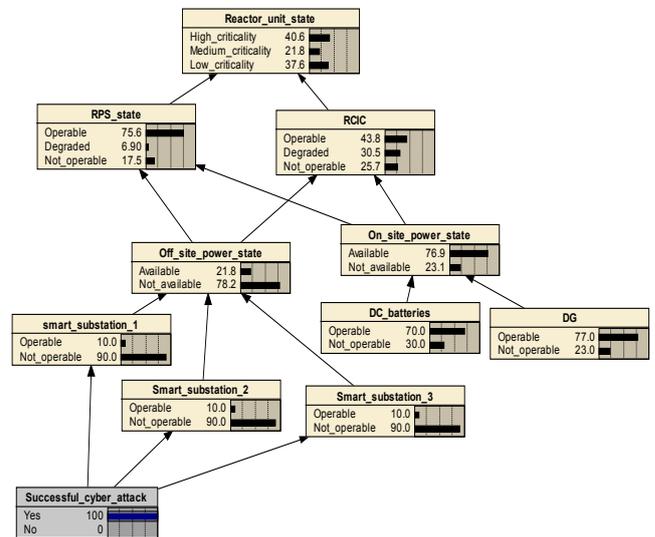


Fig. 1 BBN without cyber diversity implementation

We have selected the most different IEDs considering the approaches given above.

BBN presented below is made on parameters that consider diversity in substations. It might be seen that the probability of CCFs is decreased to 0,126.

III. CONCLUSIONS

The safe operation of NPP requires that smart substation operates in secure manner. In the future if not being treated

now the cyber risk of smart substation can compromise NPP safety. The way to assure the security of smart substation the cyber diversity is suggested in this paper.

All substations with critical load should be selected with diversity in mind. The features of diversity approach application in this case are stipulated by system of systems issues. Using cyber diversity we should take into account that different effect can be obtained for different security attributes (integrity, confidentiality, accessibility).

The approach for diversity assessment of such substations based on processing of linguistic values given by experts is suggested in this paper.

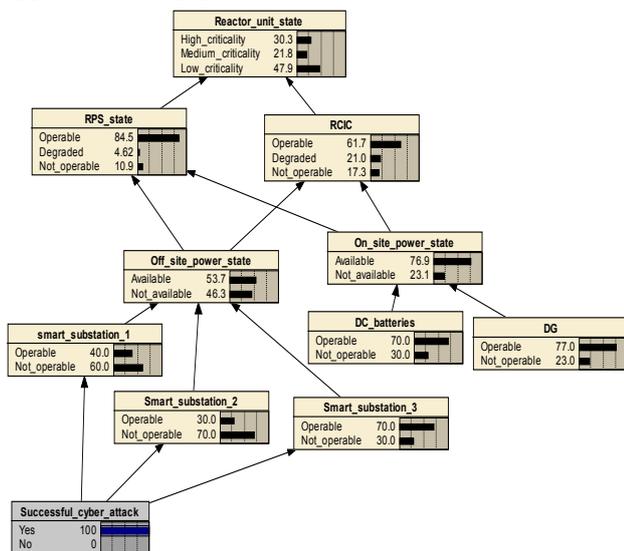


Fig. 2 BBN with cyber diversity implementation

Each IED is characterized by fuzzy diversity score of its similarity (difference) with IED that has been already selected. The cost of IED is also taken into consideration. This approach might be useful during the initial stage of substation modernization. BBN is used to evaluate CCF of substations with critical load before and after diversity implementation. Hence, diversity types and capacity selection to assure required level of security and security attributes minimizing cost is one of the future tasks.

REFERENCES

- [1] Risk management: a tool for improving Nuclear Power Plant performance, IAEA VIENNA, IAEA-TECDOC-1209, 2001.
- [2] NISTB Assessments Summary Report: Common Industrial Control System Cyber Security Weaknesses, Idaho National Laboratory Idaho Falls, Idaho 83415, 2010.
- [3] Common Cybersecurity Vulnerabilities in Industrial Control Systems, Home Land Security, Control Systems Security program, National Cyber security Division, 2011.

[4] Max Wandera, Brent Jonasson, Cybersecurity considerations for electrical distribution systems. White Paper WP152002EN, 2014.

[5] Common vulnerabilities in critical infrastructure control systems, Sandia National Laboratories Albuquerque, NM 87185-0785, 2nd edition, 2003.

[6] NPP I&C Systems for Safety and Security. M. Yastrebenetsky, V. Kharchenko (editors). USA, IGI-Global, 2014.

[7] Kharchenko V., Siora A. and Sklyar V. Multiversion FPGA-based NPP I&C Systems: Evolution of Safety // Nuclear Power Plants: Reliability, Human Factor/ N. Tsvetkov (ed.), InTech, Croatia, 121-148, 2011.

[8] Zadeh L. and Kacprzyk J. Computing with Words in Information/Intelligent Systems – Part 1: Foundation; Part 2: Applications. Heidelberg, Germany: Physica-Verlag, vol.1, 187 – 201, 1991.

Eugene Brezhnev was born in Ukraine, 1972. PhD (2000), associate professor of Computer System and Network Department, National Aerospace University “KhAI”, Kharkiv, Ukraine. His academic interest is closely related with critical infrastructure analysis, smart grid both safety and security complicated by ambiguity and incompleteness of initial data.

Vyacheslav Kharchenko was born in Ukraine, 1952. PhD (1981), Professor (1992), Doctor of Science (1995). Head of Computer Systems and Networks Department, National Aerospace University “KhAI” and Centre of Safety Infrastructure-Oriented Research and Analysis, Kharkiv, Ukraine. He is a Member of ERCIM-SERENE group, IEEE Global Education in Microelectronics Systems (I-GEMS), national supervisor of EU funding projects in the area of safety software and FPGA-based critical systems (NPP I&Cs, aerospace), green computing and communication.

Jüri Vain received the B.S. degree in system engineering from the Tallinn Polytechnic Institute, Estonia, and the Ph.D. degree from the Institute of Cybernetics at the Estonian Academy of Sciences, in 1979 and 1987, respectively. He is currently a Professor at the Department of Computer Science, Tallinn University of Technology (TUT) and he also holds a position of senior researcher at the Department of Control Systems, Institute of Cybernetics at TUT. His research interests include embedded systems, modeling of discrete-event and hybrid dynamic systems, formal verification in system design, and fault-tolerance.

Artem Boyarchuk was born in Ukraine, 1982. MSc in Computer Engineering (2005), PhD in Computer Science (2012). Research interests: methods of assessment and ensuring of SOA-based systems and infrastructures. Head of the Technology Transfer center of National Aerospace University, Kharkiv, Ukraine, is in charge of R&D cooperation with European universities and research agencies. His particular field of expertise lies in the domain of science and technology policies, innovation and technology transfer, best practice seminars.

Green computing within the context of educational and research projects

Vyacheslav Kharchenko, Oleg Illiashenko, Chris Phillips, Jüri Vain

Abstract— The paper summarizes concepts of green computing and is about the one of the most important challenges caused by a power crisis in context of information technologies application. Main definitions and taxonomy of green computing are provided. The article represents main objectives and specific tasks of TEMPUS GreenCo project and other projects, funded by European Commission – TEMPUS SEREIN, TEMPUS CABRIOLET and by FP7 Programme – KhAI-ERA. The deliverables of TEMPUS GreenCo project and connections with other TEMPUS projects are described.

Keywords — green IT, green computing, taxonomy, TEMPUS GreenCo project.

I. INTRODUCTION

Green technologies are innovations (innovative technologies) based on the principles on sustainable development and reuse of resources. Their main goal lies in decreasing of negative influence on the environment, e.g. reduce the amount of waste, increase energy efficiency, improve the design to reduce the amount of resources consumed [1]. Green technology and *green energy* is complemented by the concept *green business*, which is characterized by the use of such technologies (and such energy) at different stages of production, corporate and social responsibility, which are directly related to the concept of *green culture* [2]-[3].

During literature review and analysis on green computing and green IT, it is needed to select [4-7]. Part of them addressed problem of resources saving in IT offices, and it was not just about energy, but also other resources, including consumables (paper, toner cartridges, etc.). Among the first books of it should be noted book on green IT for the "Dummies" [8]. Later on several books (joint monographs) were published [9]-[10]. The first one should be noted specifically it more fully covers all the problems of green IT. It is interesting to note that the main sections of the book were in tune with the list of courses that have been included in the project TEMPUS GreenCo. The second book continues the series of publications on green data centers. The paper goal is to summarize concepts and taxonomy of green computing. The article represents objectives and specific tasks of TEMPUS GreenCo and other projects, funded by European Commission. The deliverables of TEMPUS GreenCo project and connections with projects TEMPUS SEREIN and CABRIOLET are described.

II. GREEN COMPUTING. BIG PICTURE

A. Green computing taxonomy

Information technologies can be described and viewed in the narrow and broad sense in the context of sustainable development. In a narrow sense – it is energy-saving and energy-efficient information technology and systems in broad one – they take into consideration environmental, security and sustainable development in general.

So, main entities in the field of sustainable development (sustainability) and green IT, as an "IT Sustainability Set (ITSS)", could be described by the Cartesian product of two subsets:

- Sustainability Development Set (SDS) of: energy and resource (EnF), ecology (EcF), safety (SF) и social and economic (CF);
- Two-element set (Means-Object Set – MOS):

$$\begin{aligned} ITSS &= \\ &= SDS * MOS = \\ &= (EnF, EcF, SF, CF) * (Means, Object) \end{aligned} \quad (1)$$

Table 1 represents this multiplication. Each cell contains particular task in the field of green IT. String "Means" isn't parted to the components because when IT is used as a tool, the combinations of components is used as a rule. Moreover strictly speaking both net and infrastructure means (tools) include hardware and software. To the strings "Object" yet another is added, which brings together all the components. This matrix allows conducting preliminary analysis of green IT directions and corresponding facts. It generalizes the similar matrices proposed by the project participants TEMPUS GreenCo [11]-[12]. The analysis performed allows developing and analyzing taxonomic scheme of green information technologies. In this article, we confine our analysis to the taxonomy of green IT in the narrow sense: define the hierarchy, clarify the basic concepts and logical connections between them.

Green energy – energy derived from renewable sources without affecting or with acceptable damage from the point of view of sustainable development for environment human and technical facilities.

Green technology – is an innovative technology, understood as a set of methods, processes and materials used in any kind

of business to create new tangible or intangible products, and based on the principles of sustainable development, obtaining and use of green energy.

SET, MOS		SETS OF FACTORS OF SUSTAINABLE DEVELOPMENT, SDS			
		EnF	EcF	SF	CF
MEANS		Means * EnF	Means * EcF	Means * SF	Means * CF
Object, IT- components	HW	EnF * HW	EcF * HW	SF * HW	CF * HW
	SW	EnF * SW	EcF * SW	SF * SW	CF * SW
	NW	EnF * NW	EcF * NW	SF * NW	CF * NW
	IS	EnF * IS	EcF * IS	SF * IS	CF * IS
	IT	EnF * IT	EcF * IT	SF * IT	CF * IT

Tab. 1. Matrix of tasks of green IT in a context of sustainability concept

Green engineering – a special kind of engineering or, as is often said, engineering, based on green technologies. It could be represented in a form of services tended to improve energy efficiency, safety and environmental performance of industrial processes and products.

Green IT – is an adaptation of the practice of IT's development and application so as to use IT more effectively. It forms a vector of development.

Green computing – a special kind of computing. In order to clarify the concept of green computing, it is necessary to understand the nature and evolution of the paradigm of computing in general. We can say that today there is a concept of computing in a narrow, broad and global sense.

In a narrow sense computing – is a calculation (the set of transformations that are performed by applying a finite number of pre-defined rules) that runs on a computer or in a computer system.

In a broad sense computing – is a collection of scientific knowledge engineering methods and activities aimed at the development and application of computer technology, including hardware and software, in different areas for the purpose of information and automation.

In a global sense, computing or noocomputing – is part of the noosphere, a theory which has developed a great scientist V.I. Vernadsky [13]. In the processes of globalization noocomputing, which formation can be completed within this decade [14], should play a key role.

Definition of **green communications** in a similar manner is projected on computer networks and telecommunications, and can be considered as a component of green computing.

Green information technology (green IT) – a set of processes, methods and tools for data collection, storage, processing, supply, distribution of information and methods of their implementation, aimed at improving energy efficiency, safety and environmental technologies themselves and the systems in which they are applied, as well as dissemination of the relevant values in the society.

Green IT engineering – kind of engineering based on the development and application of green IT in different types of human activity. Therefore, it is possible to distinguish and properly interpret the green computer engineering and green software.

Based on the concepts discussed the following components

can be determined:

- **green hardware** minimizes power consumption and the risk of dangerous failures when using systems important to safety. Accordingly, one can speak of green chips, microprocessors, modules, etc.;

- **green software** minimizes the information and energy systems, as well as based on the code, optimized on the energy metrics.

IT systems (infrastructures), based on green hardware and / or software may be called **green IT systems (infrastructures)**.

Green cloud computing can be defined as technology that can provide potential benefits for the environment and energy savings in the provision of services via the Internet or other distributed computing technologies.

Greenware – is a combination of hardware and software, and services that allow the user to minimize the effect of using a computer or computer system on the environment and the cost of expenses for their use and maintenance.

IT greening (or **greenwashing** by [9]) and **green IT reengineering** are process terms and are particularly important when it comes to developing of energy efficient software, as well as modernization of existing IT systems of different nature in the interests of reducing power consumption, etc.

Green IT culture – these are the values and norms of behavior related and directed on preservation and enhancement of all components of the environment, resources, energy and safety by improving the development and implementation of green technologies and information systems, as well as methods of professional and social activities, IT specialists in forming, development, dissemination and adoption of these values and norms.

Green IT business aims at developing and implementation of IT and green technologies, which affirmed the value of green culture, implementing a business organization that minimizes the use of energy and other resources, and direct or indirect CO2 emissions in creating products and services.

Green IT policy – is a set of goals and activities regulating the achievement of rational results in the field of green IT, specific indicators on energy efficiency and resource conservation.

B. Green computing metrics

Energy saving and energy efficiency are the main characteristics; the first one is determined by the consumption of energy, and the second – how efficiently the energy is used. It is about power, being measured by capacity P_e .

Thus energy E_e is the complex indicator that can be calculated as the ratio of capacity to be achieved, precision, and other characteristics, or their growth ΔP using technology or IT systems per watt P_e or its changes

$$\Delta P_e : E_e = \Pi / P_e \quad (2)$$

or

$$E_e = \Delta I / \Delta P_e \quad (3)$$

Energy saving has a broader meaning and indicates not only the quantitative value of energy savings by using green IT, but also on a set of measures aimed at reducing consumption.

In order to estimate the share of energy P_e , which is consumed by equipment of such system with respect to the total energy P_s of technical complex, enterprise, or any object that is embedded in the software and hardware, they use simple index MEI (power IT-system metric):

$$PIM = P_e / P_s \quad (4)$$

When talking about data centers (data center computing clusters, cloud infrastructure) one should use used metric PUE (power usage effectiveness):

$$PUE = P_s / P_e \quad (5)$$

The following metrics are also used:

GEC (green energy co-efficient) – defines the part of the energy P_{er} , which is derived from renewable sources: $GEC = P_{er} / P_e$.

ERF (energy reuse factor) determines the fraction of energy (heat primarily), released during the calculations work of and data centers as a whole, beneficial use in the future P_{er} (e.g. for heating, greenhouses, etc.):

$$EPF = P_{er} / P_e \quad (6)$$

Carbon usage effectiveness takes into account the impact of data centers and similar systems on the environment, the measured CO_2 emissions, and is determined by the ratio of emissions CE caused by a common data center power P_s to the volume of energy consumed for processing information P_e [16]:

$$CUE = CE / P_e \quad (7)$$

For the IT system, its components and processes associated with the development and application the particular indicators, which are sometimes called green metrics [17] could be calculated. They are described more detailed in [18]-[19] and in broad sense are based on so called GAMES-approach (Green Active Management of Energy in IT Service Centres). Such metrics indicate:

- the proportion of processes aimed at reducing the use of resources, including reducing energy consumption and improving energy efficiency;
- assessment of the relative influence of each of these processes and project activities on resources, energy consumption and energy efficiency;
- degree of improvement in resource characteristics of the

products obtained at different stages, etc.

III. TEMPUS AND FP7 PROJECTS ACTIVITIES

A. TEMPUS GREENCO tasks and structure

At the time of TEMPUS GreenCo project proposal there were no MSc programmes in Green IT Ukraine and other post-USSR countries. Hence it was needed to fill in the formed GAP in teaching of green IT in Academia. Concept, main activities, tasks and structure of TEMPUS GreenCo project (Green Computing and Communication), reference number 530270-TEMPUS-1-2012-1-UK-TEMPUS-JPCR, were developed at department of computer systems and networks (CSN) of National Aerospace University KhAI together with colleagues from Newcastle University. Detailed evolution of another joint TEMPUS projects is described in [20].

Specific *outcomes of TEMPUS GreenCo project are [20]:*

1. To introduce a green computing & communications programme for master students in universities in Ukraine and Russian Federation;
2. To introduce a green computing & communications programme for doctoral students in universities in Ukraine and Russian Federation;
3. To facilitate intensive capacity building measures for Ukrainian and Russian IT tutors;
4. To establish two PhD incubators in Ukraine and Russian Federation on green computing & communications.

Main activity of such TEMPUS-funded project is development of teaching courses, their dissemination for MSc, PhD and LLL (long-life learning) level. In frame of TEMPUS GreenCo project the following courses are to be developed:

1. MSc courses:

- Foundations of green IT-engineering
- Technologies of green computing
- Technologies of green regulators and robotics
- Technologies of green communication

2. PhD courses

- Standardization of green computing and communication
- Research and development (R&D) for green FPGA-based systems
- R&D for green mobile applications
- R&D for green wireless networks
- R&D for green cloud computing
- R&D of ITs for smart energy infrastructures
- Green software

3. LLL (long-life learning) courses

- Techniques and tools (T&T) for green computing
- T&T for green control systems
- T&T for green communication and management

MSc courses form a conceptual base of green ITs and provide review of existed technologies in the field of energy-saving computing, network decisions and decisions for automated process control system and robotic systems. They are intended for IT bachelor graduates to gain an understanding of the green computing methodologies and paradigms, energy efficient system level software such as compilers, hypervisors, monitoring and profiling tools, workload managers, and

programming environments, energy aware large scale distributed systems, such as Grids and Clouds. They are suitable for those aspiring to be software developers, software architecture designers, FPGA developers, experts on distributed infrastructures.

PhD courses include materials of research and development of “hot-topic” directions in green IT on the level of microcircuits, systems, networks and infrastructures. Moreover the normative regulation issues in the field of green IT and related fields are reviewed.

LLL (long-life learning) courses are aimed at practical aspects – engineering techniques and tools for development and management of green IT systems.

B. TEMPUS Greenco project deliverables

In frame of TEMPUS GreenCo two-volume multi-lecture edition covering different aspects of green IT concept was developed [21]-[22]. It presents lecture material including theoretical and practice issues of green IT-engineering for MSc-courses, PhD-Courses and training modules developed in frameworks of the project TEMPUS GreenCo.

Volume 1 [21] contains material based on the outcomes of analysis, research and development in the area of green (energy saving and energy effective) computer components and systems. The base concepts, principles and taxonomy of green computing and green IT-engineering are described. The methods and techniques of green hardware and software development and assessment for CPU- and FPGA-based embedded and mobile systems are proposed. The models and algorithms allowing save resource and power consumption at modeling, development and verification of FPGA design, high performance systems and lightweight cryptography systems are analyzed.

Volume 2 [22] contains material based on the outcomes of analysis, research and development in the area of green (energy saving) computer systems, networks, cloud IT-infrastructures and their application in industry and in context of green culture of society as a whole. The methods, technologies and cases related to the following directions are described: green Wi-Fi and mobile systems and networks (routing in networks, sensor networks, adaptation, hybrid systems); green databases and cloud computing (access and storage of data, data centers, architecting, management); green IT and industrial instrumentation and control (I&C) systems (smart grid, I&C for industry and universities); green IT for business and society (web for green, economic issues, IT-cooperation).

Figure 1 shows taxonomical scheme which describes the links between main concepts of green computing and green IT engineering and correspondent volumes of “Green IT engineering” book, published in frame of TEMPUS GreenCo project. In the left part of the scheme four groups of concepts related to the following issues are identified:

- Sustainability and its components (environmental, safety, and resource);

- Engineering, technology and systems;
- Culture, including green culture;
- Business, including green business.

On the right side, in addition to the above terms the logical connections between them are specified:

- Green computing, noocomputing;
- Green IT and Green IT engineering;
- Green IT software (components, processes, properties, characteristics and metrics) and green IT systems;
- Green IT culture and green IT business.

C. Other TEMPUS and FP7 projects

TEMPUS GreenCo project is closely connected with other projects performed at CSN department at KhAI:

SEREIN project [23] “Modernization of postgraduate studies on security and resilience for human and industry related domains”. This project is performed in frame of TEMPUS programme. To reach the main objective the international MSc and PhD programme on cyber security and resilience for Ukrainian universities will be developed. Power consumption and its leakage allow conducting successful power analysis-based attacks. Security is often about balance of protection measures and cost, so students should be able to resources efficiently.

CABRIOLET project [24] “Model-oriented approach and intelligent knowledge-based system for evolvable academia-industry cooperation in electronic and computer engineering”.

The project is aimed on design and nation-wide sustainable introduction of model-oriented approach based on implementation of general life cycle model of Evolvable Academia-to-Business Cooperation (EA2BC) and 3 customized models serving different types of Academia-to-Business cooperation and introduction of intelligent knowledge-based system (IKBS) for analysis, processing and generating of assessment results and recommendations for involved academic departments and companies. In order to establish such sustainable system the results of GreenCo project are qualitatively and effectively used.

KhAI-ERA project [25] is a special support action funded by the European Commission within FP7 Capacities Specific Programme which aims to reinforce National Aerospace University “KhAI” research cooperation capacities in order to become more closely integrated into the European Research Area (ERA). CSN department at KhAI play key role in topic C “Dependable Embedded Systems” of KhAI-ERA project (covering the topics related to architecting, development, formal methods and verification of dependable FPGA- and software-based embedded systems). Energy-saving plays in important role in embedded systems of critical application, and so the experience obtained during GreenCo project allows to act more effectively in this case.

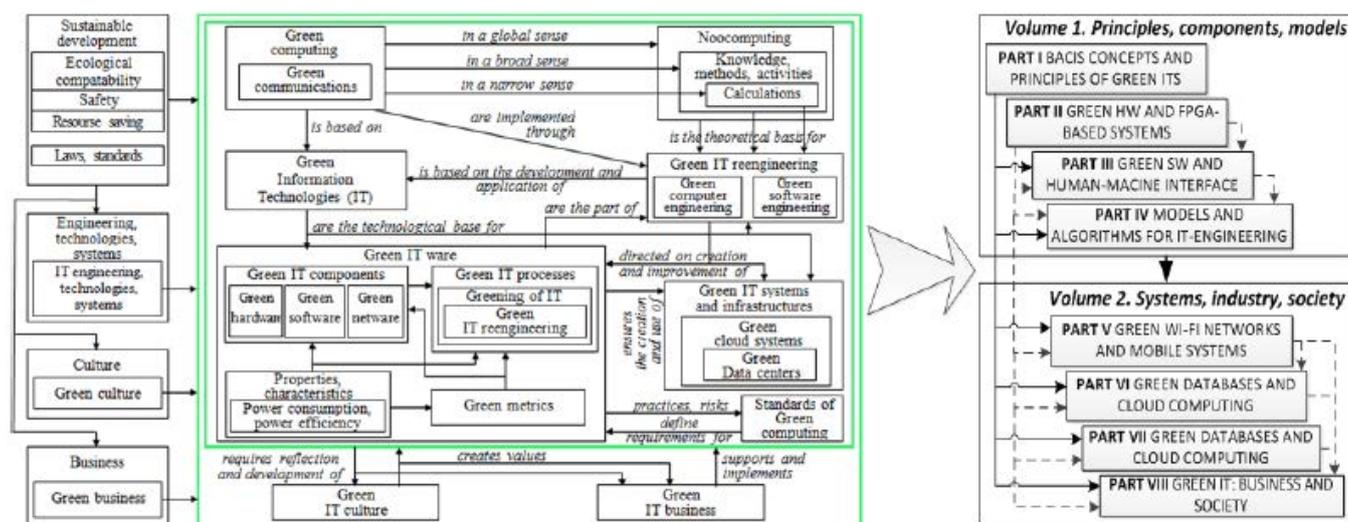


Fig. 1. Taxonomical scheme of green computing and structure of "Green IT engineering" book

KhAI's experts established permanent contacts with CEBE-TUT's [26] trainers and lecturers in order to organize training modules and deliver workshops in dependable embedded systems for experienced researchers and summer schools for young researchers. Training modules to be developed are intended for KhAI's experienced, mid-level and young participants (as well as in GreenCo project for MSc, PhD and LLL levels, SEREIN project for MSc, PhD and In-service training) selected in competitive basis taking into account their teaching activity, experience, gender issues and language skills. The overall picture of reviewed projects is described in detail in [15].

D. National green related projects

Amongst international projects the department of computer systems and networks of National Aerospace University "KhAI" took part in the national project "Theoretical foundations, methods and information technology of critical software and hardware development under resource limitations (2012-2014)". The following tasks were considered:

1. Development of scientifically grounded principles, models and methods for use of language-based approach for creating critical software systems

2. Development of principles of multiversion computing multiparameter adaptation and methods of dependability insurance for I&Cs and infrastructures taking into account resource constraints

One more project that is currently performed at CSN department of KhAI is "Scientific fundamentals, methods and tools of green computing and communications" (2015-2017).

IV. CONCLUSION

Green computing plays an important role in formation of sustainable development of human civilization. The described concepts and taxonomical scheme of green computing are a base for defining connections with other IT-domains. The

links between GreenCo, SEREIN, CABRIOLET and KhAI-ERA projects allow getting synergy effect for the teams of developers. Energy saving, green and safe ITs should be considered in computing more due to in tend to obtain sustainable society.

REFERENCES

- [1] Green technologies [Online]. Available: <http://greenevolution.ru/enc/wiki/zelenye-texnologii>
- [2] Sidorov N. A. Green information systems and technologies // Software engineering, 3(7), 2011. – pp. 5-12.
- [3] Ghauri M.R. How to go green as a telecommunication company. Master Thesis in Sustainable Development at Uppsala University, 2013. – 52 p.
- [4] Toby Velte, Anthony Velte, Robert Elsenpeter. Green IT: Reduce Your Information System's Environmental Impact While Adding to the Bottom Line, McGraw Hill Companies, 2008. – 281p.;
- [5] John Lamb. The Greening of IT: How Companies Can Make a Difference for the Environment, IBM Press, 2009. – 305p.;
- [6] Greg Schulz. The Green and Virtual Data Center Hardcover, Taylor & Francis Group, 2009. – 367p.;
- [7] Marty Poniatowski. Foundation of Green IT: Consolidation, Virtualization, Efficiency, and ROI in the Data Center, Prentice Hall, 2009. – 352 p.
- [8] Carol Baroudi, Jeffrey Hill, Arnold Reinhold, Jhana Senxian. Green IT For Dummies, Willey Publishing Inc., 2009. – 349p.
- [9] Harnessing green IT: principles and practices/ San Murugesan, G. R. Gangadharan (eds), John Wiley and Sons Ltd, 2012. – 389 p.;
- [10] The Green Computing Book/ Wu-chun Feng (edit.), Taylor & Francis Group, 2014. – 337 p.
- [11] Kharchenko V., Gorbenko A., Sklyar V., Phillips C. Green computing and communications in critical application domains: challenges and solutions // Proceeding of Digital Technologies (DT'2013), Zilina (Slovak Republic), 29-31 May 2013, pp. 191-197.
- [12] Kharchenko V., Boyarchuk A., Brezhnev E., Gorbenko A., Phillips C., Sklyar V. Green Information Technologies: The Trends in Research, Development and Education Domains Proceeding of ACSN Conference, 2013, September 16-19, Lviv, Ukraine. – 4p.
- [13] Anoprienko A., Civilization, noosphere and noorhythms // «Noosphere and civilization». Scientific journal. Issue 7 (10). – Donetsk, 2009, pp. 62-69.
- [14] Wolfengagen V., Computing: range of issues and characteristics <http://jurinfor.ru/elibs/articles/vew09s02/vew09s02.pdf>
- [15] FPGA-based critical computing: TEMPUS and FP7 projects issues / Kharchenko V., Illiashenko O., Boyarchuk A., Phillips C., Vain J., Krispin M. The 10th European workshop on microelectronics Education (EWME), 2014. – 74-79 p.

- [16] Maughan A., PUE, CUE and DCeP. Can Metrics Rescue Green IT? Morrison & Foerster LLP, 15 April 2014. – P.1-5
- [17] <http://www.theguardian.com/environment/2014/apr/02/social-media-explosion-powered-dirty-coal-greenpeace-report>.
- [18] Christian Belady (ed.). Carbon Usage Effectiveness (CUE): A Green Grid Data Center Sustainability Metric, The Green Grid, 2010. – 8p.
- [19] Kipp A., Tao Jiang, Fugini M. Green Metrics for Energy-aware IT Systems Complex, Intelligent and Software Intensive Systems (CISIS), International Conference on Date of Conference, June 30-July 2 2011. – P. 241-248.
- [20] TEMPUS GreenCo project website [Online]. Available: <http://my-greenco.eu/>
- [21] Green IT-Engineering. One-volume edition, Vol.1. Principles, components models. / Kharchenko V. (edit) – Department of Education and Science of Ukraine, National aerospace university “KhAI”. - 2014. - 594 p.
- [22] Green IT-Engineering. One-volume edition, Vol.2. Systems, industry, society. / Kharchenko V. (edit) – Department of Education and Science of Ukraine, National aerospace university “KhAI”. - 2014. - 688 p.
- [23] TEMPUS SEREIN project web-site [Online]. Available: <http://serein.net.ua/>
- [24] TEMPUS CABRIOLET project web-site [Online]. Available: <http://www.my-cabriolet.eu/>
- [25] FP7 KhAI-ERA project web-site [Online]. Available: <http://khai-era.khai.edu/>
- [26] Centre for Integrated Electronic Systems and Biomedical Engineering – CEBE web-site [Online]. Available: <http://khai-era.khai.edu/http://cebe.ttu.ee/>

Vyacheslav Kharchenko was born in Ukraine, 1952. PhD (1981), Professor (1992), Doctor of Science (1995). Head of Computer Systems and Networks Department, National Aerospace University “KhAI” and Centre of Safety Infrastructure-Oriented Research and Analysis, Kharkiv, Ukraine. He is a Member of ERCIM-SERENE group, IEEE Global Education in Microelectronics Systems (I-GEMS), national supervisor of EU funding projects in the area of safety software and FPGA-based critical systems (NPP I&Cs, aerospace), green computing and communication.

Oleg Illiashenko was born in Ukraine, 1989. MSc in Computer Engineering (2012) and MSc in Information Security (2013). Assistant lecturer of Computer Systems and Networks Department, National Aerospace University “KhAI”. Information manager of TEMPUS GREENCO, CABRIOLET, SEREIN projects at National Aerospace University “KhAI”. Research interests: safety and security assessment, assurance and regulatory aspects of critical I&C systems.

Chris Phillips currently Dean of Undergraduate Studies in the Faculty of Science, Agriculture and Engineering but continue to have teaching involvement as a Senior Lecturer in the School of Computing Science. Chris joined Newcastle in 1984 after working for five years in the Computing Studies Department at Hull University, and before that in the Statistics and Computational Mathematics Department at Liverpool University, where he also gained his BSc, MSc and PhD. Chris' background is as a numerical analyst/computational scientist. Research interests are in the area of pedagogic research.

Jüri Vain received the B.S. degree in system engineering from the Tallinn Polytechnic Institute, Estonia, and the Ph.D. degree from the Institute of Cybernetics at the Estonian Academy of Sciences, in 1979 and 1987, respectively. He is currently a Professor at the Department of Computer Science, Tallinn University of Technology (TUT) and he also holds a position of senior researcher at the Department of Control Systems, Institute of Cybernetics at TUT. His research interests include embedded systems, modeling of discrete-event and hybrid dynamic systems, formal verification in system design, and fault-tolerance.

Evolution of software quality models: usability, security and greenness issues

Oleksandr Gordieiev, Vyacheslav Kharchenko and Mario Fusani

Abstract – software quality models (SQMs) are usually represented by a set of interconnected characteristics (subcharacteristics). SQMs were analyzed beginning with first SQMs. At that time (from NATO conference in 1968 year [1]) «program engineering» was based as independent direction at IT. Some characteristics in SQMs are competitive, i.e. increase of one characteristic lead to deterioration to another characteristic. The analysis presented in this paper results from noticing an evolution of competitive SQM characteristics, such as security and usability. Another emerging aspect, i.e. (sub)characteristics regarding sustainable software, or “green” software (greenness) is investigated as well. Mutual influence and comparative analysis of security, usability and greenness are described using a set of metrics considering relevance of characteristics in SQMs and their changing during last 40 years.

Keywords — software quality model, software greenness, software security, software usability, evolution analysis, metrics, ISO/IEC9126, ISO/IEC25010, structure-semantic analysis.

I. INTRODUCTION

Results of applying Structure-Semantic Analysis (SSA) technique [2] to SQM analysis give us the possibility from one side to observe changes in SQMs during more than 40 years of software engineering, and from another side, to determine development trends for each SQM characteristic separately [3]. In our previous works we proposed metrics for assessing relevance of SQM characteristics, subcharacteristics and models as a whole. Was established, that change of nomenclature and structure of characteristics such as security, usability, characteristics which associated with green and subcharacteristics, which itemize of them. Changes in SQMs are determined by trends in «software engineering» technologies development. Preliminary analysis of derived results in [2-5] allowed authors of the paper assume, that changes in competitive characteristics SQMs interconnected among themselves. In article [3] among between reliability and greenness has been confirmed.

Goal of the paper is to analyse evolution for some competitive characteristics of SQMs, such as a security and usability. Besides, subcharacteristics regarding green software (greenness) are researched as well. Mutual influence and comparative analysis of security, usability and greenness are described using a set of metrics considering relevance of characteristics in SQMs and their changing during last 40 years.

II. CHOICE OF COMPETITIVE CHARACTERISTICS

Preliminary analysis of SQMs allowed determined potential most competitive characteristics. We selected such characteristics - usability, security and greenness.

Different aspects of attributes relation are described for pairs “usability-security” [6,7], “security-greenness” [8,9], “usability-greenness” [10,11]. Analysis of the works allowed concluding about interesting triangle which is formed by these characteristics and evolution during last decades.

Greenness characteristics are not contained in existing SQMs in an explicit form. Therefore green related or associated with software greenness characteristics and subcharacteristics are analyzed in the paper. Farther, characteristics and characteristics are by SQM structure-semantic analysis (SSA) technique [2].

III. SSA-TECHNIQUE DESCRIPTION

Let us shortly describe the sequence of SSA-technique and to use for analysis of competitive characteristics. The technique describes quality models as a facet-hierarchy structure (graph). Nodes correspond to quality attributes and links take into account hierarchy dependencies. To briefly characterize the proposed technique of analysis let us introduce some initial terms: conceptual model is a model which a model under study is compared with; model under study is a model which is compared with a conceptual model; characteristic under study is a conceptual model characteristic which is compared with model under study characteristics.

SSA-technique is based on comparing a model under study with the conceptual model, i.e. every SQM is compared with the conceptual model. So, the analysis is equivalent to semantically comparing characteristics and subcharacteristics of a model under study and the conceptual model with regard to their structures. Selecting a reference model is usually performed by an expert who has relevant experience and qualifications.

At the following stage comparison of models among themselves should be performed. The simplest and most obvious metrics are offered. Hierarchy of these metrics is presented in Fig. 1. The metrics are used to compare models with reference model bottom up, i.e. first at the level of subcharacteristics (subcharacteristics matching metric SMM, cumulative subcharacteristics comparison metric CSCM, characteristics matching metric CMM), then at the level of characteristics (cumulative matching characteristics metric CMCM) and finally at the level of models as a whole

(cumulative software quality models comparison metric CSQMCM).

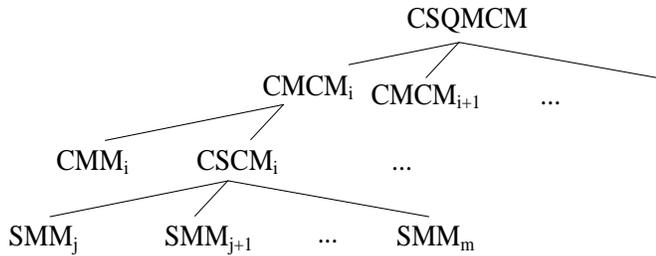


Fig. 1 Metrics hierarchy

Features of the metrics are the following:

- subcharacteristic matching metric (SMMj). Every subcharacteristic match value is identified as SMMj = 0,5 / number of reference (conceptual) model elements subcharacteristics of the characteristic under study. Weights of characteristics are not considered when calculating metrics;

- cumulative subcharacteristics comparison metric (CSCM) is evaluated as a sum of SMM:

$$CSCM_i = \sum_{j=1}^k SMM_j; \quad (1)$$

- characteristics matching metric (CMM) takes the value of 0.5 in case of matching or 0 if the characteristics are different;

- cumulative matching characteristics metric (CMCM) is calculated as a sum:

$$CMCM_i = CMM_i + \sum_{j=1}^k CSCM_j; \quad (2)$$

- cumulative software quality models comparison metric (CSQMCM) is calculated according to the formula:

$$CSQMCM_i = \sum_{j=1}^n CMCM_j \quad (3)$$

IV. RESULTS OF EVOLUTION ANALYSIS AND INTERFERENCE OF COMPETITIVE CHARACTERISTICS

Let us conduct SW QM analysis and first of all, define the reference (conceptual) model. SW Quality Model ISO/IEC 25010 [12] will be considered as uppermost and etalon regarding to all other models. It is the newest introduced model and takes into account main modern software peculiarities from the point of view of quality evaluation. This model is described by international standard of top level.

According with results of analysis CMCM is calculated for the set of characteristics presented in Table 1. The results of calculation are shown in Table 2 (Chs – characteristics, SChs – subcharacteristics) for Greenness, Usability and Security characteristics.

The histogram of CMCM values for software quality models is presented in Fig. 2 (black color for Greenness, gray for Usability and light gray for Security). An abscissa axis corresponds to years of SQM emergence. Initial point (year) is 1970 (as a first year after 1968 which is multiple of ten years).

Table 1. SWQM characteristics (greenness, security, usability)

№	SWQMs (years)	Greenness characteristics	Security characteristics	Usability characteristics
1.	McCall (1977)	4. Efficiency	-	6. Usability
		4.1 Execution efficiency		6.1 Operability
		4.2 Storage efficiency		6.2 Training
2.	Boehm (1978)	2.2 Efficiency	-	3.2 Understability
		2.2.1 Accountability		3.2.1 Legibility
		2.2.2 Accessibility		3.2.2 Conciseness
				3.2.3 Structureness
				3.2.4 Self descriptiveness
3.	Carlo Ghezzi (1991)	-	-	7. Usability
4.	FURPS (1992)	4. Performance	1. Functionality	2. Usability
		4.1 Velocity	1.3 Security	2.1 Human factors
		4.2 Efficiency		2.2 Aesthetic
		4.3 Availability		2.3 Documentation of the user
		4.4 Time of answer		2.4 Material of training
		4.5 Time of recovery		
		4.6 Utilization of resources		
4.6 Capacity				
5.	IEEE (1993)	1. Efficiency	1. Functionality	6. Usability
		1.1 Temporal efficiency	1.3 Security	6.1 Comprehensibility
		1.2 Resource efficiency		6.2 Ease of learning
			6.3 Communicativeness	
6.	Dromey (1995)	2.2 Efficiency	-	-
7.	ISO 9126-1 (2001)	4. Efficiency	1. Functionality	3 Usability
		4.1 Time behavior	1.4 Security	3.1 Understandability
		4.2 Resource utilization		3.2 Learnability
			3.3 Operability	
				3.4 Attractiveness
8.	QMOOD (2002)	6 Effectiveness	-	3. Understandability
9.	ISO 25010 (2010)	2. Performance efficiency	6. Security	4. Usability
		2.1 Time behavior	6.1 Confidentiality	4.1 Appropriateness recognisability
		2.2 Resource utilization	6.2 Integrity	4.2 Learnability
		2.3 Capacity	6.3 Non-repudiation	4.3 Operability
			6.4 Accountability	4.4 User error protection
			6.5 Authenticity	4.5 User interface aesthetics
			4.6 Accessibility	

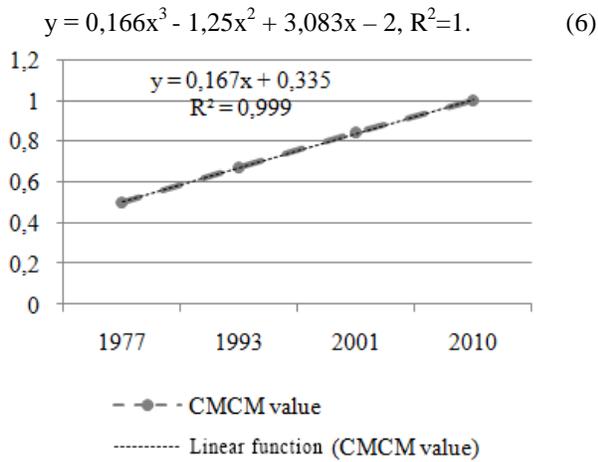


Fig. 4 Diagram of change of CMCM values for Greenness

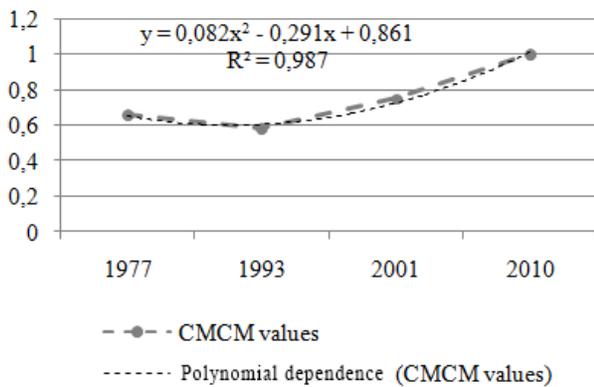


Fig. 5 Diagram of change of CMCM values for Usability

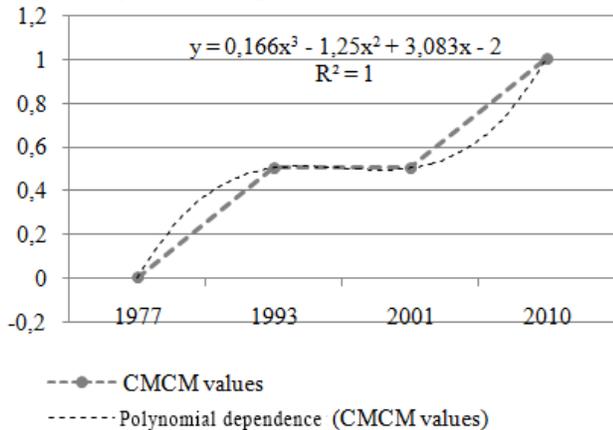


Fig. 6 Diagram of change of CMCM values for Security

Formulas 4-6 and figures 4-6 illustrate a tendency of SQMs characteristics/ subcharacteristics changes. Analysis of dependencies (Fig. 3) allows concluding that weights of Greenness, Usability and Security characteristics became equal in 2010 (the standard ISO/IEC 25010).

Each competitive characteristic was described by different types of mathematical relations. Exact choice of relation type was confirmed high value of coefficient of determination (R^2).

Derived results gave us following information:

- difference between minimal and maximal values of metric CMCM (Δ CMCM) determine dynamic of evolution for

each characteristic: for Greenness Δ CMCM = 0,5; for Usability Δ CMCM = 0,34; for Security Δ CMCM = 1;

- modifications of CMCM metric values for Security characteristic in SWQM give us the basis to approve that this characteristic has undertaken dynamic development recently. Firstly, as subcharacteristic, security was represented in 1992 year in SWQM FURPS, and as single characteristic only in 2010 year in SWQM ISO 25010. In this perspective, we can confirm that evolution of Security characteristic is developing with delay in SWQM;

- CMCM metric value for Greenness and Security characteristics during their evolution evolution did not decrease, and for Usability slightly decrease by 0,8 in 1993 in IEEE SQM;

- we can confirm, that linear mathematical relation, which was obtained for Greenness characteristic, reveals natural development of Green Software technologies;

- nonlinearity of mathematical relation for Usability, to a lesser degree, and for Security farther give us information, that such type of relation was determined by important external factors. For example, for Security such factor is growth of number intrusion to information systems for insufficient quality of software Security.

Influence of Security on another characteristics (Usability and Greenness) was determined. For this interaction influence of sub characteristics was determined as shown in Table 3.

Table 3. Result of subcharacteristics interaction

Security	Usability			Greenness					
	Appropriateness recognisability	Learnability	Operability	User error protection	User interface aesthetics	Accessibility	Time behavior	Resource utilization	Capacity
Confidentiality	~	~	~	↑	↓	↓	↑	↑	↓
Integrity	~	↓	↓	~	↓	~	↑	↑	↓
Non-repudiation	~	~	~	~	↓	~	↑	↑	↓
Accountability	~	~	~	~	↓	~	↑	↑	↓
Authenticity	~	~	~	↑	↓	↓	↑	↑	↓

Table of symbols:
 ↓ worsening of one subcharacteristic in the presence of improvement of another characteristic (security);
 - worsening of one subcharacteristic in the presence of improvement of another characteristic (security) do not occur ;
 ~ unknown dependence.

V. CONCLUSION

In result of this work more competitive characteristics of SQMs, such as Greenness, Usability and Security were selected. For each of these characteristics analysis was conducted and interaction was determined.

Security characteristic from three competitive characteristics is developing more dynamically. Such tendency is represented by sharp increase of software security requirements recently.

Authors consider that in modern conditions SQMs should change more often than 1 time in 10 years. More often changes in SQMs can be connected with only separated characteristics, for example, with security.

REFERENCES

- [1] NATO SCIENCE COMMITTEE Report. Software engineering. Report on a conference sponsored by the NATO SCIENCE COMMITTEE, 136 p., Germany, Garmisch
- [2] Gordieiev O., Kharchenko V., Fominykh N., Sklyar V. Evolution of software quality models in context of the standard ISO 25010. Proceedings of 9th International Conference on Dependability and Complex Systems (DepCoS-RELCOMEX 2014), 30 Jun-4 July, 2014, Advances in Intelligent and Soft Computing, Springer, Brunow, Poland, pp. 223-233,
- [3] Gordieiev O., Kharchenko V., Fusani M. Evolution of Software Quality Models: Green and Reliability Issues. Proceedings of the 11th International Conference ICT in Education, Research and Industrial Applications: Integration, Harmonization and Knowledge Transfer (ICTERI 2015), May 14, 2015 Lviv, Ukraine: CEUR-WS.org, pp. 432-445.
- [4] Biscoglio I., Coco A, Fusani M., Gnesi S., Trentanni G., "An approach to Ambiguity Analysis in Safety-related Standards", In Proceedings of QUATIC 2010, Porto, Portugal, pp. 461-466.
- [5] Ferrari A., Fantechi A., Gnesi S., Magnani G. Model-based development and formal methods in the railway industry. In: IEEE Software, vol. 30 (3) IEEE, 2013, pp. 28-34.
- [6] Bryan D. Payne, W. Keith Edwards. A Brief Introduction to Usable Security, Useful Computer Security, May/June, 2008. pp. 13-21.
- [7] Dirk Balfanz, Glenn Durfee, D.K. Smetters. In Search of Usable Security: Five Lessons from the Field, IEEE Security & Privacy, September/October, 2004, pp. 19-24.
- [8] Xun Li, Frederic T. Chong. A Case for Energy-Aware Security Mechanisms. Proceedings of 27th International Conference Networking and Applications Workshops (WAINA 2013), 25-28 March, 2013, Barcelona, Spain, pp. 1541-1546.
- [9] Carroll M., Merwe A., Kotze P. Secure cloud computing: Benefits, risks and controls. Proceedings of Information Security South Africa (ISSA 2011), 15-17 Aug., 2011, IEEE, Johannesburg, South Africa, pp. 1-9.
- [10] Effie Lai-Chong Law, Ebba Thora Hvannberg, Gilbert Cockton. Maturing Usability. Part 6 A Green Paper on Usability Maturation. Human-Computer Interaction Series. Springer-Verlag London, 2008, pp. 381-424.
- [11] Thimbleby H. Interaction Walkthrough: Evaluation of Safety Critical Interactive Systems. The XIII International Workshop on Design, Specification and Verification of Interactive Systems (DSVIS 2006), Springer Lecture Notes in Computer Science, 2007, pp. 52-66.
- [12] International Standard ISO/IEC 25010. Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models, ISO/IEC JTC1/SC7/WG6.

Oleksandr Gordieiev was born in Ukraine, 1981. PhD (2007), associate professor (2010), associate professor of security banking management Department, University of banking of the National bank of Ukraine. His academic interest is closely related with software quality (include assessment), green computing and communication, usability assessment.

Vyacheslav Kharchenko was born in Ukraine, 1952. PhD (1981), Professor (1992), Doctor of Science (1995). Head of Computer Systems and Networks Department, National Airspace University "KhAI" and Centre of Safety Infrastructure-Oriented Research and Analysis, Kharkiv, Ukraine. He is a Member of ERCIM-SERENE group, IEEE Global Education in Microelectronics Systems (I-GEMS), national supervisor of EU funding projects in the area of safety software and FPGA-based critical systems (NPP I&Cs, aerospace), green computing and communication.

Mario Fusani, PhD, has been with the Systems & Software Evaluation Centre, ISTI (Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo") – CNR (National Research Council), Pisa, Italy, since 1973. He participated in projects on standard development (ISO, IEC and CENELEC), quality models for standards, standard analysis, researched natural-language requirements analysis and disambiguation. Research interests are standardization of requirements and certification of critical software and safety systems for railway, automotive and other critical domains.

Simulation on Friction Welding Of MgAZ31 / AA 6061 T6 Joints

N.Rajesh Jesudoss Hynes and P.Shenbaga Velu

Abstract:In the recent years, leading car manufacturers are exploring the possibility of joining magnesium to aluminium for diverse automotive applications. Friction welding is the most suitable candidate for joining of dissimilar materials in all these critical applications. In the present work, numerical studies were carried out to study the mechanism of joining of MgAZ31 alloy with Al 6061T6 alloy by Friction welding. The developed thermo-mechanical model is highly non-linear due to the interaction between temperature field and time dependent material properties. It could be used as an industrial tool to predict the nodal temperature, stress, deformation and heat flux of the dissimilar joints.

Keywords:Friction welding, Al 6061 T6 Alloy, MgAZ31 Alloy, Temperature distribution, Deformation.

1. Introduction

Friction welding is a solid state joining process that uses the rotational

motion and high axial pressure to convert rotational energy into frictional heat a circular interface. The basic principle of friction welding is one of the components being welded is rotated while the other is kept stationary. The two components are then brought together by an axially applied force. Rubbing the two surfaces together produces sufficient heat in such a way that local plastic zones are formed and axially applied force causes the plasticised metal to be extruded from the joint, carrying with contaminants, oxides etc.

1.1 Literature Survey:

Hazman Seli et al [1] performed numerical analysis of friction welding of Mild steel / Al 6061 T6 alloy combinations. All the experiments performed were modelled using ABAQUS software. Akbari Mousavi [2] and Rahbar Kelishami performed numerical analysis of friction welding of 4340 steel/ Mild steel combinations .Wenya Li et al [3] modelled the 2 dimensional friction welding of dissimilar joints at various conditions and to improve the results 3 dimensional model were developed using ABAQUS software. Jolanta Zimmerman [4] developed the

N.Rajesh Jesudoss Hynes is with Mepco Schlenk Engineering College, Virudhunagar, Tamil Nadu, India (e-mail: findhynes@yahoo.co.in).

P.Shenbaga Velu is with Mepco Schlenk Engineering College, Virudhunagar, Tamil Nadu, India

thermo mechanical model of ceramic & metal friction welding of dissimilar joints using FEM software. Rajesh et al [5] performed numerical simulation on joining of Ceramics with Metal by Friction Welding technique.

2. Numerical Modelling

Friction welding is a complicated metallurgical process that is accompanied by frictional heat generation and plastic deformation. Friction welding Al 6061 T6 alloy & MgAZ31 alloy was numerically simulated using ABAQUS Software. In the present work, the simulation is to predict the Temperature distribution, stress, deformation, strain and strain rate during the joining process.

2.1 FEM Model:

2.1.1 Assumptions:

The assumptions made when defining the loads and boundary conditions are as follows,

1. 100% of dissipated energy caused during the friction between parts was converted to heat and distributed evenly between two interfacing surfaces.
2. The cylindrical rods were assumed to experience frictional contact described by Coulombs frictional law with temperature dependent friction coefficient μ .

2.1.2 Boundary Conditions:

The boundary conditions apply convection mode of heat transfer for the external surfaces and conduction at conduct surfaces of Al 6061 T6 alloy-MgAZ31 alloy (Fig 2). Initial temperature of the room is assumed to be 29 °C.

2.1.3 Mesh and Geometry:

The Al 6061 T6 alloy & MgAZ31 alloy rods are modelled individually with 10 mm diameter and 20 mm length by using ABAQUS Software. The Element type of the dissimilar material is C3D8RT which has an 8-node thermally coupled brick structure with tri-linear displacement. (Fig 1)

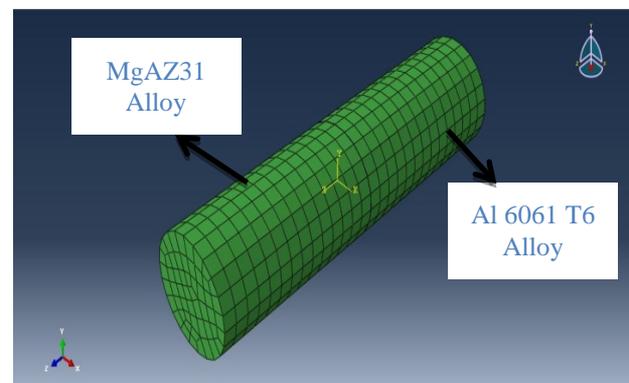


Fig.1. Meshed model

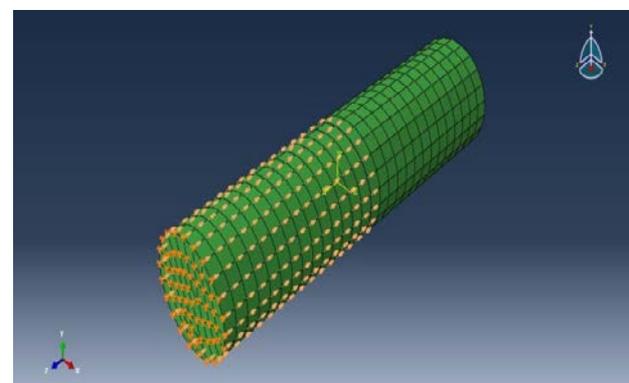


Fig.2. Applying loading and boundary conditions

2.1.4 Thermo Mechanical Model:

The thermal model was sequentially coupled to mechanical model. The

mechanical loading of stationary rod is taken into account while retaining the load step size used in the thermal model. The temperature history of the rod was considered in each load step with mechanical loading to predict the stress developed in the work pieces.

The numerical simulation of the friction welding process was carried out with MgAZ31 alloy rod was rotating at an angular velocity of 93.4 rad/sec. The predicted heat flux $1.85e6 \text{ W/m}^2$ is applied to both the interface surfaces. The simulation was carried out to predict the heat flux vector, nodal temperature, stress, strain and deformation of the dissimilar materials joined by Friction welding process.

3. Coupled Field Analysis:

(Al 6061 T6 Alloy Vs Mg AZ31 Alloy)

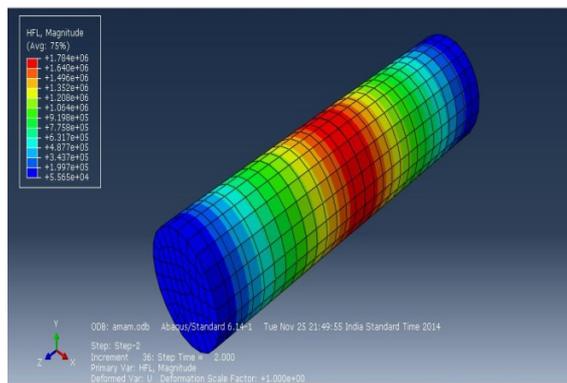


Fig.3. Heat Flux Vector

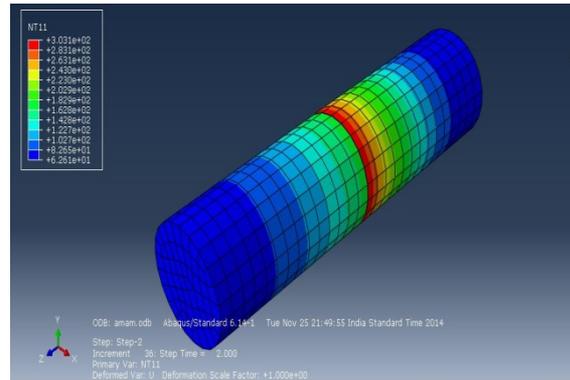


Fig.4. Nodal Temperature

The temperature is very high at the joint interface of the dissimilar materials and the gradient of the temperature is gradually decreases in the axial direction. Peak temperature reaches $303.1 \text{ }^{\circ}\text{C}$ at the heating of during fiction time (Fig 4).The heating temperature reaches 75% of the melting point of the material. The rate of heat transfer towards the end of both the material varies due to different thermal properties of the respective material. (Fig 3)

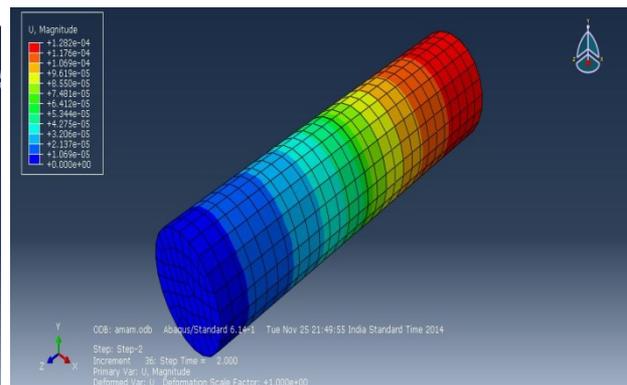


Fig.5. Displacement of dissimilar materials

During thewelding process deformation occurs near to the interface of the dissimilar joint. At every moment of the friction phase the plasticized material is expelled out of the interface due to the visco plastic flow of the materials. (Fig 5)

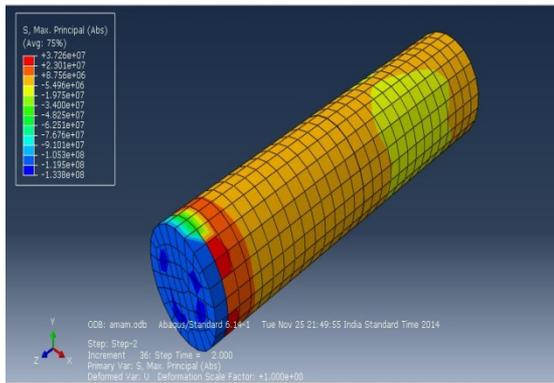


Fig.6. Maximum principal Stress

Figure 6 results shows the maximum principal stress occurs away from a distance from the weld interface. When the temperature increases, the yield stress decreases which in turn reduce the von mises stress at the interface of the dissimilar joint.

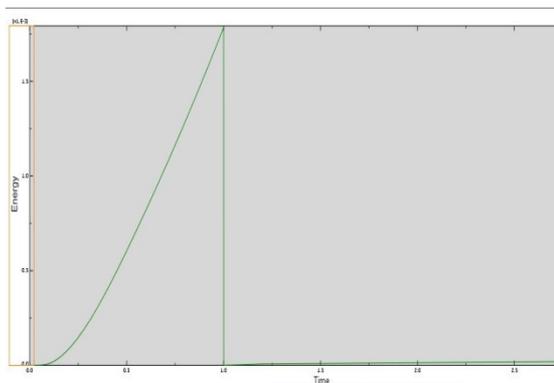


Fig.7.Contact Constant stain energy of the Whole model

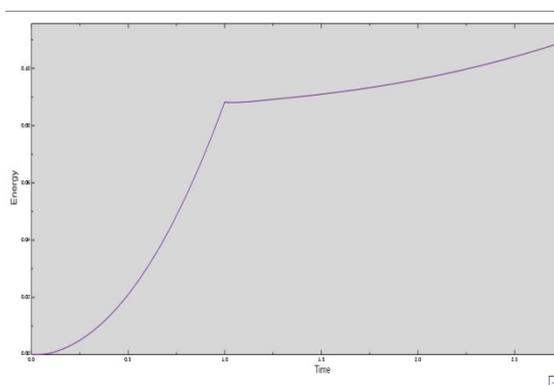


Fig.8. Total energy of the output for whole model

Figure 7 shows the results of constant contact strain energy for whole model, at initial time of the welding process the strain energy is very high once the process is completed the strain energy is reduced gradually decreased to attain the constant value. The total energy of the output for whole model (Fig 8) is increases while increase the time period for the friction welding of dissimilar joint.

4. CONCLUSION:

Numerical simulations were performed to analyze the friction welding process that involves high temperature, large deformation and transient operations. The coupled effects of the mechanical and heat transfer are taken into account in the numerical model. The distributions of temperature, heat flux vector, displacement and stress during the friction welding were numerically predicted.

The following conclusions are drawn from this study.

1. The peak von mises stress is produced at distance of 3.5 mm away from the centerlines.
2. The highest temperature is obtained at the weld centerline and the temperature decreases axially away from the centerline.
3. During joining of Al 6061 T6 alloy with MgZ31 alloy, the temperature distribution is found to be lower than in the MgZ31 Alloy side due to change in thermal properties.

4. The energy increases with the increase in time for the whole strain energy model.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support of this work by SERB of Department of Science & Technology, New Delhi. (Vide Letter No.: SERB / F / 1452 / 2013-2014 dated 10.06.2013).

REFERENCES

1. Hazman Seli et al, 'Evaluation of Properties and FEM Model of the Friction Welded Mild Steel-Al6061-Alumina', Materials Research.16 (2): 453-467, 2013.
2. Wenya Li et al, 'Numerical Simulation of Friction Welding Processes Based on ABAQUS Environment' Journal of Engineering Science and Technology Review 5 (3) 10-19, 2012.
3. Jolanta Zimmerman et al, 'Thermo Mechanical and diffusion modeling in the process of ceramic-metal friction welding' Journal of materials processing and technology 209, 1644-1653, 2009.
4. Rajesh et al, 'Numerical Simulation on Joining of Ceramics with Metal by Friction Welding Technique', International Journal of Modern Physics: Conference Series, 22, 190-195
5. Mohamad Zaky Noha, LuayBakirHussainb, Zainal Arifin Ahmadb, 'Alumina-mild steel friction welded at lower rotational speed' Journal of materials processing technology 204, 279-283, 2008.
6. V.V. Satyanarayana a, G. MadhusudhanReddyb, T. Mohandasb 'Dissimilar metal friction welding of austenitic-ferritic stainless steels' Journal of Materials Processing Technology 160, 128-137, 2005.
7. R. Paventhan , P.R. Lakshminarayanan , V. Balasubramanian 'Fatigue behaviour of friction welded medium carbon steel and austenitic stainless steel dissimilar joints' Materials and Design 32, 1888-1894, 2011.
8. S.D. Meshram, T. Mohandas, G. Madhusudhan Reddy 'Friction welding of dissimilar pure metals' Journal of Materials Processing Technology 184, 330-337, 2007.
9. Radosław Winiczenko, MieczysławKaczorowski 'Friction welding of ductile cast iron using interlayers' Materials and Design 34 , 444-451, 2012.
10. L.D'Alvise, E.Massoni, S.J.Walloe 'Finite element modeling of the inertia friction welding process between dissimilar material' Journal of Materials and Processing Technology 125-126, 387-391, 2002.
11. Rajesh et al, 'Finite Element Based Thermal Modeling of Friction Welding of Dissimilar Materials', International Journal of Modern Physics: Conference Series, 22, 96-202.
12. Rajesh et al, 'Numerical simulation of heat flow in Friction Stud Welding of dissimilar metals', Arabian Journal for Science and Engineering, 39, 3217-3224, 2014.

Authors Index

Abelha, A.	51, 354, 366	Daineko, Y. A.	340	Imeraj, D.	201
Adamkó, A.	44	Dangat, S. S.	483	Ipalakova, M. T.	340
Adamski, M.	281	Dascal, I.	215	Iracleous, D. P.	63
Aher, K. V.	457	De Paolis, L. T.	243	Isidro, R. L.	388
Ahn, G.	167	Dho, A. Y.-H.	502	Jacob, B.	497
Ajhoun, R.	238	Dias, E. M.	259, 267	Jašek, R.	170
Akbarizadeh, G.	469	Dias, M. L. R. P.	267	Jelonek, D.	29, 249
Alaoui, S.	238	Dmitriyev, V. G.	340	Jeong, J.	206
Alavi, S. E.	469	Dontas, E.	476	Jestin, V. K.	440
Alejandro, R. V.	388	Doukas, N.	63, 476	Jesudoss Hynes, N. R.	524
Ally, M.	34	Dulík, T.	170	Ji, J.	163
Andruseac, G. G.	463	Dvořák, V.	115	Jia, W.	418
Apostolidis, H.	103	El Hajje, M.-J.	292	Jiang, B.	57, 97
Arturo, P. P.	388	El Idrissi, Y. El B.	238	Jiang, K.	418
Atanasov, I. I.	394	Elfattah, M. M. A.	406	Jones, I.	281
Bae, M.	163	Fernandes, S. L.	427, 434	Jung, S.	163
Bala, G. J.	427, 434	Figueiredo, A. E. P.	259	Kachá, P.	298
Belyaev, K. P.	192	Figueiredo, L.	259	Kádek, T.	44
Biadacz, R.	151	Frankovic, A.	227	Kalou, Ai. K.	360
Blišňák, M.	170	Fritz, A.	281	Khair, M.	292
Bogdanova, A. I.	314	Fusani, M.	519	Khalil, L.	292
Borkovec, R.	40	Galaction, A.-I.	463	Kharchenko, V.	507, 513, 519
Borza, S.	254	Galka, D.	281	Kierzynka, M.	281
Botsios, S. D.	360	Giyenko, A. D.	340	Kim, D.	163
Bourro, K. E.	63	Glavan, L. M.	227	Kim, I.	206
Boyarchuk, A.	507	Goliński, M.	184	Kim, J.	163
Brezhniev, E.	507	Gomes, J.	348	Kim, Yon.	163, 167
Bubel, D.	177	Gordieiev, O.	519	Kim, You.	163
Calderwood, M.	75	Grover, P. S.	371	Kirsininkas, R. J.	92
Carvalho, R. de D.	259	Halili, A.	201	Kiselev, A. B.	314
Chae, H.	286	Halilovic, A.	249	Kollár, L.	44
Chandar, K. P.	308, 335	Hamiti, M.	400	Kolouch, J.	304
Chauhan, S. K.	371	Hassan, S.	488	Kósa, M.	44
Chen, Y.	75	Hoeflich, S.	259	Kostěnc, M.	298
Cheptea, C.	463	Holman, D.	139	Koutsomitropoulos, D. A.	360
Churbanova, N. G.	197	Hong, H.	167	Krawczyk-Sokołowska, I.	145
Ciobanu, M.	215	Hong, K.	286	Kropáčová, A.	298, 304
Coufal, P.	40	Hurezeanu, B.	414	Kuleshov, A. A.	192, 197
Da Silva, E. B.	267	Hwang, J.	167	Lee, C.	163
Daaboul, T.	292	Illiashenko, O.	513	Lee, J.	167

Lee, K.-W.	502	Petránek, K.	211	Tanajura, C. A. S.	192
Lenort, R.	139	Phillips, C.	513	Taralunga, D. D.	414
Li, J.	57, 75	Ploscar, A.	215	Theocharis, J. B.	327
Lim, H.	163, 167	Pluhacek, M.	121	Topaloglou, C. A.	327
Lim, Y.	163, 167	Portela, F.	51, 348, 354	Torres, L.	354
Lima, L.	366	Portela, F.	366	Toussaint, G. T.	86
Łobaziewicz, M.	131	Poștaru, M.	463	Trapeznikova, M. A.	197
Łukasik, K.	145	Qerimi, E.	201	Tsiatsos, T.	103
Lyupa, A. A.	197	Radnaeva, V. D.	234	Tuchkova, N. P.	192
Maalem, S.	449	Rakhimzhanova, N. K.	340	Tyurenkova, V. V.	314
Machado, J.	51, 354, 366	Rexha, B.	201	Ungureanu, M.	414
Macura, L.	376, 382	Rezac, F.	376, 382	Vain, J.	507, 513
Maier, D.	281	Ricciardi, F.	243	Vardhana, M.	445
Manes, C. L.	243	Robisch, P.	92	Varela, D. T.	81
Mastorakis, N. E.	68	Rozhon, J.	376, 382	Velu, P. S.	524
Mastorocostas, P. A.	327	Safarik, J.	376, 382	Virag, I.	215
Mehenni, T.	275	Salem, A.-B. M.	343	Vlahovic, N.	219, 227, 320
Mekhaznia, T.	108	Santos, M. F.	51, 348, 354	Voznak, M.	376, 382
Milková, E.	211	Santos, M. F.	366	Vukšić, V. B.	320
Mirkazemy, A.	469	Savithri, T. S.	308, 335	Wang, L.	97
Mohamad, T. Z.	343	Sedivy, J.	40	Wang, M.	57, 75, 92
Mostafa, A. H.	406	Senkerik, R.	121	Wang, M.	97, 103
Mylonas, S. K.	327	Shah, S.	488	Waykar, S. B.	457, 497
Naaji, A.	215	Shalbuev, D. V.	234	Wells, A.	281
Nagavci, D.	400	Sharma, A.	371	Wicher, P.	139
Neves, J.	354	Simion, C.	254	Włodarczak, P.	34
Nikitin, V. F.	314	Slachta, J.	376, 382	Wyslocka, E.	151
Novosád, J.	115	Smirnov, N. N.	314	Xiao, J.	57, 97
Oliveira, P.	51	Smirnova, M. N.	314	Yang, Y.	167
Oplatkova, Z. K.	121	Soar, J.	34	Youssif, A. A. A.	406
Pánovics, J.	44	Solomou, G. D.	360	Zhang, C.	418
Park, E.	167	Staš, D.	139	Zharnikova, E. V.	234
Patil, S. R.	483	Stavrakoudis, D. G.	327	Zhuang, X.	68
Pencheva, E. N.	394	Stępnia, C.	157	Zidani, A.	108
Pereira, S. L.	259, 267	Strungaru, R.	414	Ziora, L.	127
Pestov, D. A.	314	Szafranski, M.	184	Zoroja, J.	320