ADVANCES in INFORMATION SCIENCE and APPLICATIONS - VOLUME II

Proceedings of the 18th International Conference on Computers (part of CSCC '14)

**Santorini Island, Greece
July 17-21, 2014**

ADVANCES in INFORMATION SCIENCE and APPLICATIONS - VOLUME II

Proceedings of the 18th International Conference on Computers (part of CSCC '14)

Santorini Island, Greece
July 17-21, 2014

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 | 23

ISSN: 1790-5109
ISBN: 978-1-61804-237-8

ADVANCES in INFORMATION SCIENCE and APPLICATIONS - VOLUME II

Proceedings of the 18th International Conference on Computers
(part of CSCC '14)

Santorini Island, Greece
July 17-21, 2014

Organizing Committee

Editors:
Prof. Nikos Mastorakis, Technical University of Sofia, Bulgaria and HNA, Greece
Prof. Kleanthis Psarris, The City University of New York, USA
Prof. George Vachtsevanos, Georgia Institute of Technology, Atlanta, Georgia, USA
Prof. Philippe Dondon, École Nationale Supérieure d'Électronique, Talence, Cedex, France
Prof. Valeri Mladenov, Technical University of Sofia, Bulgaria
Prof. Aida Bulucea, University of Craiova, Craiova, Romania
Prof. Imre Rudas, Obuda University, Budapest, Hungary
Prof. Olga Martin, Politehnica University of Bucharest, Romania

Associate Editors:
Antoanela Naaji
Abdel-Badeeh M. Salem
Elena Zamiatina
Luca De Cicco
Antonio Pietrabissa

Steering Committee:
Prof. Theodore B. Trafalis, University of Oklahoma, USA
Prof. Charles A. Long, Professor Emeritus, University of Wisconsin, Stevens Point, Wisconsin, USA
Prof. Maria Isabel García-Planas, Universitat Politècnica de Catalunya, Spain
Prof. Reinhard Neck, Klagenfurt University, Klagenfurt, Austria
Prof. Myriam Lazard, Institut Superieur d' Ingenierie de la Conception, Saint Die, France
Prof. Zoran Bojkovic, University of Belgrade, Serbia
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. George Vachtsevanos (Georgia Institute of Technology, 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. Yuriy S. Shmaliy, (IEEE Fellow, The University of Guanajuato, Mexico)

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. Panos Pardalos (University of Florida, USA)
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. Tadeusz Kaczorek (IEEE Fellow, Warsaw University of Tehcnology, 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)
Prof. S. Furfari (Direction Generale Energie et Transports, Brussels, EU)
Prof. Constantin Udriste, University Politehnica of Bucharest , ROMANIA
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. Photios Anninos (Medical School of Thrace, Greece)
Prof. Ichiro Hagiwara, (Tokyo Institute of Technology, Japan)
Prof. Andris Buikis (Latvian Academy of Science. Latvia)
Prof. Akshai Aggarwal (University of Windsor, Canada)
Prof. George Vachtsevanos (Georgia Institute of Technology, USA)
Prof. Ulrich Albrecht (Auburn University, USA)
Prof. Imre J. Rudas (Obuda University, Hungary)
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. Elias C. Aifantis (Aristotle Univ. of Thessaloniki, Greece)
Prof. Anastasios Lyrintzis (Purdue University, USA)
Prof. Charles Long (Prof. Emeritus University of Wisconsin, USA)
Prof. Marvin Goldstein (NASA Glenn Research Center, USA)
Prof. Costin Cepisca (University POLITEHNICA of Bucharest, Romania)
Prof. Kleanthis Psarris (University of Texas at San Antonio, USA)
Prof. Ron Goldman (Rice University, USA)
Prof. Ioannis A. Kakadiaris (University of Houston, USA)
Prof. Richard Tapia (Rice University, USA)
Prof. F.-K. Benra (University of Duisburg-Essen, Germany)
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. Elbrous M. Jafarov (Istanbul Technical University, Turkey)
Prof. M. Ehsani (Texas A&M University, USA)
Prof. Sesh Commuri (University of Oklahoma, USA)
Prof. Nicolas Galanis (Universite de Sherbrooke, Canada)
Prof. S. H. Sohrab (Northwestern University, USA)
Prof. Rui J. P. de Figueiredo (University of California, USA)
Prof. Valeri Mladenov (Technical University of Sofia, Bulgaria)
Prof. Hiroshi Sakaki (Meisei University, Tokyo, Japan)
Prof. Zoran S. Bojkovic (Technical University of Belgrade, Serbia)
Prof. K. D. Klaes, (Head of the EPS Support Science Team in the MET Division at EUMETSAT, France)
Prof. Kazuhiko Tsuda (University of Tsukuba, Tokyo, Japan)
Prof. Milan Stork (University of West Bohemia , Czech Republic)
Prof. C. G. Helmis (University of Athens, Greece)
Prof. Lajos Barna (Budapest University of Technology and Economics, Hungary)
Prof. Nobuoki Mano (Meisei University, Tokyo, Japan)
Prof. Nobuo Nakajima (The University of Electro-Communications, Tokyo, Japan)
Prof. Victor-Emil Neagoe (Polytechnic University of Bucharest, Romania)
Prof. P. Vanderstraeten (Brussels Institute for Environmental Management, Belgium)
Prof. Annaliese Bischoff (University of Massachusetts, Amherst, USA)
Prof. Virgil Tiponut (Politehnica University of Timisoara, Romania)
Prof. Andrei Kolyshkin (Riga Technical University, Latvia)
Prof. Fumiaki Imado (Shinshu University, Japan)
Prof. Sotirios G. Ziavras (New Jersey Institute of Technology, USA)
Prof. Constantin Volosencu (Politehnica University of Timisoara, Romania)
Prof. Marc A. Rosen (University of Ontario Institute of Technology, Canada)
Prof. Thomas M. Gatton (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. Eugene Kindler (University of Ostrava, Czech Republic)
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. Arpad A. Fay (University of Miskolc, Hungary)
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. Paul Dan Cristea (University "Politehnica" of Bucharest, Romania)
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. Ray Hefferlin (Southern Adventist University, TN, USA)
Prof. Gabriella Bognar (University of Miskolc, Hungary)
Prof. Hamid Abachi (Monash University, Australia)
Prof. Karlheinz Spindler (Fachhochschule Wiesbaden, Germany)
Prof. Josef Boercsoek (Universitat Kassel, Germany)
Prof. Eyad H. Abed (University of Maryland, Maryland, USA)
Prof. F. Castanie (TeSA, Toulouse, France)
Prof. Robert K. L. Gay (Nanyang Technological University, Singapore)
Prof. Andrzej Ordys (Kingston University, UK)
Prof. Harris Catrakis (Univ of California Irvine, USA)
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. Ion Chryssoverghi (National Technical University of Athens, Greece)
Prof. Dr. Aydin Akan (Istanbul University, Turkey)
Prof. Sarka Necasova (Academy of Sciences, Prague, Czech Republic)
Prof. C. D. Memos (National Technical University of Athens, Greece)
Prof. S. Y. Chen, (Zhejiang University of Technology, China and University of Hamburg, Germany)
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. Dr-Eng. Christian Bouquegneau (Faculty Polytechnique de Mons, Belgium)
Prof. Wladyslaw Mielczarski (Technical University of Lodz, Poland)
Prof. Ibrahim Hassan (Concordia University, Montreal, Quebec, Canada)
Prof. Stavros J.Baloyannis (Medical School, Aristotle University of Thessaloniki, Greece)
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. Walter Dosch (University of Luebeck, Germany)
Prof. Rudolf Freund (Vienna University of Technology, Austria)
Prof. Erich Schmidt (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. Vladimir Damgov (Bulgarian Academy of Sciences, Bulgaria)
Prof. P.Borne (Ecole Central de Lille, France)

Additional Reviewers

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

Table of Contents

Plenary Lecture 1

Floating Offshore Wind Turbines: The Technologies and the Economics

Prof. Paul D. Sclavounos
Professor of Mehanical Engineering and Naval Architecture
Massachusetts Institute of Technology (MIT)
77 Massachusetts Avenue
Cambridge MA 02139-4307
USA
E-mail: pauls@mit.edu

Abstract: Wind is a vast, renewable and clean energy source that stands to be a key contributor to the world energy mix in the coming decades. The horizontal axis three-bladed wind turbine is a mature technology and onshore wind farms are cost competitive with coal fired power plants equipped with carbon sequestration technologies and in many parts of the world with natural gas fired power plants.

Offshore wind energy is the next frontier. Vast sea areas with higher and steadier wind speeds are available for the development of offshore wind farms that offer several advantages. Visual, noise and flicker impacts are mitigated when the wind turbines are sited at a distance from the coastline. A new generation of 6-10MW wind turbines with diameters exceeding 160m have been developed for the offshore environment. They can be fully assembled at a coastal facility and installed by a low cost float-out operation. Floater technologies are being developed for the support of multi-megawatt turbines in waters of moderate to large depth, drawing upon developments by the offshore oil & gas industry.

The state of development of the offshore wind energy sector will be discussed. The floating offshore wind turbine technology will be reviewed drawing upon research carried out at MIT since the turn of the 21st century. Floating wind turbine installations worldwide and planned future developments will be presented. The economics of floating offshore wind farms will be addressed along with the investment metrics that must be met for the development of large scale floating offshore wind power plants.

Brief Biography of the Speaker: Paul D. Sclavounos is Professor of Mechanical Engineering and Naval Architecture at the Massachusetts Institute of Technology. His research interests focus upon the marine hydrodynamics of ships, offshore platforms and floating wind turbines. The state-of-the-art computer programs SWAN and SML developed from his research have been widely adopted by the maritime, offshore oil & gas, and wind energy industries. His research

activities also include studies of the economics, valuation and risk management of assets in the crude oil, natural gas, shipping and wind energy sectors. He was the Georg Weinblum Memorial Lecturer in 2010-2011 and the Keynote Lecturer at the Offshore Mechanics and Arctic Engineering Conference in 2013. He is a member of the Board of the North American Committee of Det Norske Veritas since 1997, a member of the Advisory Committee of the US Navy Tempest program since 2006 and a member of the Advisory Board of the Norwegian Center for Offshore Wind Energy Technology since 2009. He has consulted widely for the US Government, shipping, offshore, yachting and energy industries.
http://meche.mit.edu/people/?id=76

Keynote Lecture 2

Detecting Critical Elements in Large Networks

Professor Panos M. Pardalos
Center for Applied Optimization (CAO)
Department of Industrial and Systems Engineering,
University of Florida, Gainesville, FL, USA.
and
Laboratory of Algorithms and Technologies for Networks Analysis (LATNA)
National Research University, Higher School of Economics
Moscow, Russia
E-mail: p.m.pardalos@gmail.com

Abstract: In network analysis, the problem of detecting subsets of elements important to the connectivity of a network (i.e., critical elements) has become a fundamental task over the last few years. Identifying the nodes, arcs, paths, clusters, cliques, etc., that are responsible for network cohesion can be crucial for studying many fundamental properties of a network. Depending on the context, finding these elements can help to analyze structural characteristics such as, attack tolerance, robustness, and vulnerability. Furthermore we can classify critical elements based on their centrality, prestige, reputation and can determine dominant clusters and partitions.

From the point of view of robustness and vulnerability analysis, evaluating how well a network will perform under certain disruptive events plays a vital role in the design and operation of such a network. To detect vulnerability issues, it is of particular importance to analyze how well connected a network will remain after a disruptive event takes place, destroying or impairing a set of its elements. The main goal is to identify the set of critical elements that must be protected or reinforced in order to mitigate the negative impact that the absence of such elements may produce in the network. Applications are typically found in homeland security, energy grid, evacuation planning, immunization strategies, financial networks, biological networks, and transportation.

From the member-classification perspective, identifying members with a high reputation and influential power within a social network could be of great importance when designing a marketing strategy. Positioning a product, spreading a rumor, or developing a campaign against drugs and alcohol abuse may have a great impact over society if the strategy is properly targeted among the most influential and recognized members of a community. The recent emergence of social networks such as Facebook, Twitter, LinkedIn, etc. provide countless applications for problems of critical-element detection.

In addition, determining dominant cliques or clusters over different industries and markets via critical clique detection may be crucial in the analysis of market share concentrations and debt

concentrations, spotting possible collusive actions or even helping to prevent future economic crises.

This presentation surveys some of the recent advances for solving these kinds of problems including heuristics, mathematical programming, dynamic programming, approximation algorithms, and simulation approaches. We also summarize some applications that can be found in the literature and present further motivation for the use of these methodologies for network analysis in a broader context.

Brief Biography of the Speaker: Panos M. Pardalos serves as Distinguished Professor of Industrial and Systems Engineering at the University of Florida. He is also an affiliated faculty member of the Computer and Information Science Department, the Hellenic Studies Center, and the Biomedical Engineering Program. He is also the Director of the Center for Applied Optimization. Dr. Pardalos is a world leading expert in global and combinatorial optimization. His recent research interests include network design problems, optimization in telecommunications, e-commerce, data mining, biomedical applications, and massive computing.

Full CV: http://www.ise.ufl.edu/pardalos/files/2011/08/CV_Dec13.pdf

Recent Achievments: http://www.eng.ufl.edu/news/first-engineering-chair-appointed-under-ufs-preeminence-initiative-goes-to-big-data-expert/

Profile in Scholar Google: scholar.google.com/scholar?q=P+Pardalos&btnG=&hl=en&as_sdt=0,5

Plenary Lecture 3

Overview of the Main Metaheuristics used for the Optimization of Complex Systems

Professor Pierre Borne
Co-author: Mohamd Benrejeb
Ecole Centrale de Lille
France
E-mail: pierre.borne@ec-lille.fr

Abstract: For complex systems such as in planning and scheduling optimization, the complexity which corresponds usually to hard combinational optimization prevents the implementation of exact solving methodologies which could not give the optimal solution in finite time. It is the reason why engineers prefer to use metaheuristics which are able to produce good solutions in a reasonable computation time. Two types of metaheuristics are presented here:
* The local searchs, such as: Tabu Search, Simulated Annealing, GRASP method, Hill Climbing, Tunnelling...
* The global methods which look for a family of solutions such as: Genetic or Evolutionary Algorithms, Ant Colony Optimization, Particle Swarm Optimization, Bees algorithm, Firefly algorithm, Bat algorithm, Harmony search....

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 "Ingénieur 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 4

Minimum Energy Control of Fractional Positive Electrical Circuits

Professor Tadeusz Kaczorek (Fellow IEEE)
Warsaw University of Technology
Poland

Abstract: The talk will consist of two parts. In the first part the minimum energy control of standard positive electrical circuits will be discussed and in the second part the similar problem for fractional positive electrical circuits. Necessary and sufficient conditions for the positivity and reachability of electrical circuits composed of resistors, coils and capacitors will be established. The minimum energy control problem for the standard and fractional positive electrical circuits will be formulated and solved. Procedures for computation of the optimal inputs and minimal values of the performance indeces will be given and illustrated by examples of electrical circuits.

Brief Biography of the Speaker: Prof. Tadeusz Kaczorek graduated from the Faculty of Electrical Engineering Warsaw University of Technology in 1956, where in 1962 he defended his doctoral thesis. In 1964, he received a postdoctoral degree. In the years 1965-1970 he was head of the Department of Electronics and Automation, 1969-1970, and Dean of the Faculty of Electrical Engineering University of Warsaw. In the years 1970-1973 Vice-Rector of the Technical University of Warsaw in the years 1970-1981 the director of the Institute of Control and Industrial Electronics Warsaw University of Technology. He was also head of the Department of Control of the above Institute. In 1971 he received the title of Professor and Associate Professor of Warsaw University of Technology. In 1974 he received the title of professor of Warsaw University of Technology. In 1987-1988 he was chairman of the Committee for Automation and Robotics. Since 1986, corresponding member, and since 1998 member of the Polish Academy of Sciences. In 1988-1991 he was Head of the Scientific Academy in Rome. For many years a member of the Foundation for Polish Science. From June 1999 ordinary member of the Academy of Engineering. He is currently a professor at the Faculty of Electrical Engineering of Bialystok and Warsaw University of Technology. Since 1991 he is a member, and now chairman of the Central Commission for Academic Degrees and Titles (Vice-President in 2003-2006). In 2012 he was chairman of the Presidium of the Scientific Committee of the conference devoted to research crash of the Polish Tu-154 in Smolensk methods of science.
Scientific achievements
His research interests relate to automation, control theory and electrical engineering, including analysis and synthesis of circuits and systems with parameters determined and random polynomial methods for the synthesis of control systems and singular systems. Author of 20 books and monographs and over 700 articles and papers in major international journals such as

IEEE Transactions on Automatic Control, Multidimensional Systems and Signal Processing, International Journal of Control, Systems Science and Electrical Engineering Canadian Journal.

He organized and presided over 60 scientific sessions at international conferences, and was a member of about 30 scientific committees. He has lectured at over 20 universities in the United States, Japan, Canada and Europe as a visiting professor. He supervised more than 60 doctoral dissertations completed and reviewed many doctoral theses and dissertations. His dozens of alumni received the title of professor in Poland or abroad.

He is a member of editorial boards of journals such as International Journal of Multidimensional Systems and Signal Processing, Foundations of Computing and Decision Sciences, Archives of Control Sciences. From 1 April 1997, is the editor of the Bulletin of the Academy of Technical Sciences.

Honours, awards and honorary doctorates.

Honours

Tadeusz Kaczorek has been honored with the following awards:

* Officer's Cross of the Order of Polonia Restituta Polish

* Meritorious Polish

* Medal of the National Education Commission

Honorary doctorates

He received honorary degrees from the following universities:

Silesian University of Technology (2014)
Rzeszow University of Technology (2012)
Poznan University of Technology (2011)
Opole University of Technology (2009)
Technical University of Lodz (3 December 2008)
Bialystok University of Technology (August 20, 2008)
Warsaw University of Technology (22 December 2004)
Szczecin University of Technology (November 8, 2004)
Lublin University of Technology (13 May 2004)
University of Zielona Gora (27 November 2002)

Honorary Member of the Hungarian Academy of Sciences and the Polish Society of Theoretical and Applied Electrical (1999). He received 12 awards of the Minister of National Education of all levels (including 2 team).

Plenary Lecture 5

Unmanned Systems for Civilian Operations

Professor George Vachtsevanos
Professor Emeritus
Georgia Institute of Technology
USA
E-mail: george.vachtsevanos@ece.gatech.edu

Abstract: In this plenary talk we will introduce fundamental concepts of unmanned systems (Unmanned Aerial Vehicles and Unmanned Ground Vehicles) and their emerging utility in civilian operations. We will discuss a framework for multiple UAVs tasked to perform forrest fire detection and prevention operations. A ground station with appropriate equipment and personnel functions as the support and coordination center providing critical information to fire fighter as derived from the UAVs. The intent is to locate a swarm of vehicles over a designated area and report at the earliest the presence of such fire precursors as smoke, etc. the UAVs are equipped with appropriate sensors, computing and communications in order to execute these surveillance tasks accurately and robustly. Meteorological sensors monitor wind velocity, temperature and other relevant parameters. The UAV observations are augmented, when appropriate, with satellite data, observation towers and human information sources. Other application domains of both aerial and ground unmanned systems refer to rescue operations, damage surveillance and support for areas subjected to earthquakes and other natural disasters, border patrol, agricultural applications, traffic control, among others.

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 6

Iterative Extended UFIR Filtering in Applications to Mobile Robot Indoor Localization

Professor Yuriy S. Shmaliy
Department of Electronics
DICIS, Guanajuato University
Salamanca, 36855, Mexico
E-mail: shmaliy@ugto.mx

Abstract: A novel iterative extended unbiased FIR (EFIR) filtering algorithm is discussed to solve suboptimally the nonlinear estimation problem. Unlike the Kalman filter, the EFIR filtering algorithm completely ignores the noise statistics, but requires an optimal horizon of N points in order for the estimate to be suboptimal. The optimal horizon can be specialized via measurements with much smaller efforts and cost than for the noise statistics required by EKF. Overall, EFIR filtering is more successful in accuracy and more robust than EKF under the uncertain conditions. Extensive investigations of the approach are conducted in applications to localization of mobile robot via triangulation and in radio frequency identification tag grids. Better performance of the EFIR filter is demonstrated in a comparison with the EKF. It is also shown that divergence in EKF is not only due to large nonlinearities and large noise as stated by the Kalman filter theory, but also due to errors in the noise covariances ignored by EFIR filter.

Brief Biography of the Speaker: Dr. Yuriy S. Shmaliy is a full professor in Electrical Engineering of the Universidad de Guanajuato, Mexico, since 1999. He received the B.S., M.S., and Ph.D. degrees in 1974, 1976 and 1982, respectively, from the Kharkiv Aviation Institute, Ukraine. In 1992 he received the Dr.Sc. (technical) degree from the Soviet Union Government. In March 1985, he joined the Kharkiv Military University. He serves as full professor beginning in 1986 and has a Certificate of Professor from the Ukrainian Government in 1993. In 1993, he founded and, by 2001, had been a director of the Scientific Center “Sichron” (Kharkiv, Ukraine) working in the field of precise time and frequency. His books Continuous-Time Signals (2006) and Continuous-Time Systems (2007) were published by Springer, New York. His book GPS-based Optimal FIR Filtering of Clock Models (2009) was published by Nova Science Publ., New York. He also edited a book Probability: Interpretation, Theory and Applications (Nova Science Publ., New York, 2012) and contributed to several books with invited chapters. Dr. Shmaliy has authored more than 300 Journal and Conference papers and 80 patents. He is IEEE Fellow; was rewarded a title, Honorary Radio Engineer of the USSR, in 1991; and was listed in Outstanding People of the 20th Century, Cambridge, England in 1999. He is currently an Associate Editor for Recent Patents on Space Technology. He serves on the Editorial Boards of several International Journals and is a member of the Organizing and Program Committees of various Int. Symposia. His current interests include statistical signal processing, optimal estimation, and stochastic system theory.

PART II

The influence of the parameter h in Homotopy analysis method for boundary value problems

Wang Zhen, , Qin Yu Peng, , Zou Li

Abstract— **In this paper, we pay more attention to the embedding parameter h, which has an influence on the convergence region of solution series in the Homotopy analysis method(HAM). We use some theorems to give the concrete influence and proof. Then introduce a new modified method of the HAM called the blocked homotopy analysis method. Futhermore, examples such as NLS equation, ricatti equation and Duffing equation are presented to illustrate the main results.**

Homotopy analysis method; the convergence region; Cauchy's estimate; Cauchy-Kowalevskaya theorem;

I. Introduction

In 1992, Liao employed the basic ideas of the homotopy in topology to propose a general analytic method for nonlinear problems, namely Homotopy analysis method (HAM)[1]. Based on homotopy of topology, the validity of the HAM is independent of whether or not there exist small parameters in the considered equation. Therefore, the HAM can overcome the foregoing restrictions and limitations of perturbation techniques. This method has been successfully applied to solve many types of nonlinear problems. Absolutely all related article note that the convergence region of solution series depend upon the value of the embedding parameter h in the Homotopy analysis method. Unlike all previous analytic techniques, we can adjust and control the convergence region of solution series by assigning h a proper value. The closer the value of h is to zero, the larger the convergence region. However, we usually know it as one common things without giving a strict proof.

Here, we pay more attention to h to show the detail influence and verify it on the solution of the NLS equation with cubic nonlinearity which plays an important role in describing the slow modulation of a carrier wave in a dispersive medium.

II. Mathematical formulation

Consider the abstract boundary value problem

$$\begin{cases}(1-q)\mathcal{L}[\Phi(t,q)-u_0(t)] = hqH(t)[L(\Phi(t,q)) - N(\Phi(t,q))], t>0,\\ u_0(0) = f.\end{cases} \tag{2.1}$$

where the operator $\mathcal{L}$, L is linear and N is nonlinear, N(u) is analytic near the initial data f. Without loss of generality, we take $\mathcal{L} = L = \frac{\partial}{\partial t}, u_0(t) = f, H(t) = -1$, the part of $t > 0$ in the abstract formulation (2.1) could be transformed into

$$(1-q)\frac{\partial}{\partial t}\Phi(t,q) + hq[\frac{\partial}{\partial t}\Phi(t,q) - N(\Phi(t,q))] = 0.$$

After simplifying the above formulation, it is clear that

$$\frac{\partial}{\partial t}\Phi(t,q) = mN(\Phi(t,q)), \tag{2.2}$$

where $m := \frac{hq}{hq-q+1}$. Moreover, Eq (2.2) could be written as

$$\Phi(t,q) = f + m\int_0^t N(\Phi(s,q))ds. \tag{2.3}$$

with $\Phi(0,q) = f$.

III. Convergence analysis

lemma (Cauchy's estimate). Suppose that $\Phi(t)$ is differentiable in $\{t : |t-t_0| < T\}$, and to any $\hat{t} \in (0,T)$, there exist a $M > 0$, such that $|\Phi(t)| \le M$ on $C : |t-t_0| = \hat{t}$, then

$$|\Phi^{(k)}(t_0)| \le \frac{Mk!}{\hat{t}^k}, \forall k \ge 0.$$

proof. From the Cauchy integral formula, we obtain

$$\Phi^{(k)}(t_0) = \frac{k!}{2\pi i}\int_C \frac{\Phi(t)}{(t-t_0)^{k+1}}dt,$$

then

$$|\Phi^{(k)}(t_0)| = \frac{k!}{2\pi}|\int_C \frac{\Phi(t)}{(t-t_0)^{k+1}}dt| \le \frac{k!}{2\pi}\frac{M}{\hat{t}^{k+1}}2\pi\hat{t} = \frac{Mk!}{\hat{t}^k}$$

Theorem 1 (Cauchy-Kowalevskaya). Suppose $\Phi(t)$ is the exact solution of the abstract formulation (2.2), $m < +\infty$, then there exists a $\tau > 0$ such that $u : [0,\tau] \to R$ is also an analytic real function.

proof. As N(u) is analytic near f, $m < +\infty$, so $m^2 N(\Phi)$ is also analysis near f. Next, by Cauchy's estimate, there exist $a, b > 0$ such that

$$m^2\frac{1}{k!}|\partial_\Phi^k N(f)| \le \frac{b}{a^k}, \forall k \ge 0, \tag{3.1}$$

where $\partial_\Phi^k N(\Phi)$ make the sense of the Fréchet derivative, it means that $\partial_\Phi N(\Phi) = N(\Phi), \partial_\Phi N^2(\Phi) = N^{(2)}(\Phi)$, ... $N(\Phi)'$s taylor series at f converges to r:=$\forall|\Phi - f| < a$, and what's more, we obtain that

$$m|N(\Phi)| \le m\sum_{k=0}^{\infty}\frac{1}{k!}|\partial_\Phi^k N(f)||(\Phi-f)^k| \le \frac{b}{m}\sum_{k=0}^{\infty}\frac{|\Phi-f|^k}{a^k}$$

$$\le \frac{a\frac{b}{m}}{a-r} =: mg(r).$$

Wang zhen is with the School of Mathematical Science, Dalian University of Technology, Dalian, 116085 China e-mail: wangzhen@dlut.edu.cn.

Qin Yu Peng is with Dalian University of Technology, too.

Zou Li is with Department of Naval Architecture, Dalian University of Technology, Dalian 116085, China, emaillizou@dlut.edu.cn

Manuscript received June 29, 2014; revised July 2, 2014.

ISBN: 978-1-61804-237-8

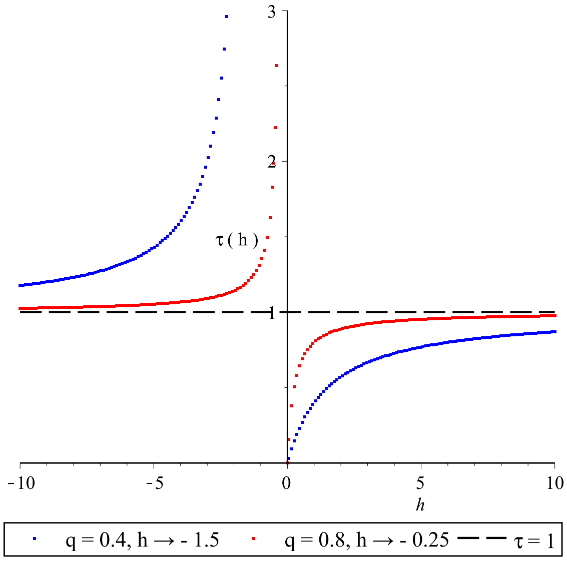


Fig. 1. Take different value of q=0.4, 0.8 with a=2,b=1, we find the closer the corresponding value of $h \longrightarrow 1-\frac{1}{q}$ is to -1.5, -0.25, the closer τ is to ∞.

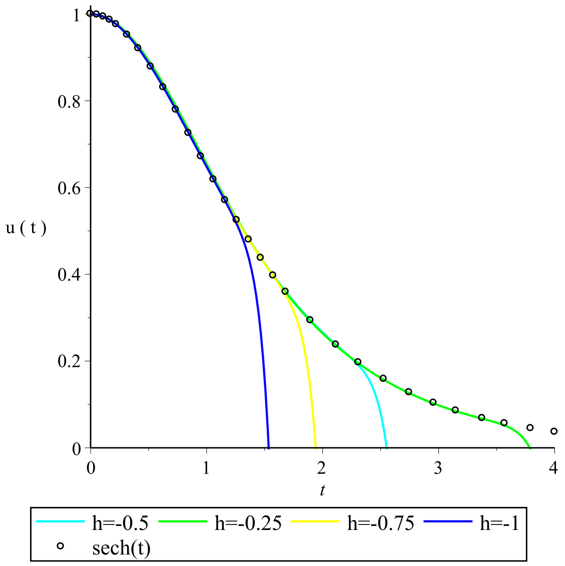


Fig. 2. Take different value of h=-1, -0.75, -0.5, -0.25, we find the value of h is closer to 0, the corresponding graph is more fit with the graph of the exact solution.

To the majorant function $g(r)$, it shows

$$m\frac{1}{k!}|\partial_\Phi^k N(f)| \le m\frac{1}{k!}\partial_r^k g(0), \tag{3.2}$$

so we can consider the majorant problem

$$\begin{cases} \dot{r}(t) = mg(r), t > 0, \\ r(0) = 0, \end{cases} \tag{3.3}$$

where $r \in \mathbb{R}_+$ and it's exact solution is

$$r(t) = a - \sqrt{a^2 - 2a\frac{b}{m}t}.$$

It is clear that $r(t)$ is analytic on $(-\infty, \frac{a}{2b}m)$. According to the comparison principle, if $\Phi(t)$ suit (2.7), then

$|\Phi(t) - f| \le m\int_0^t |N(\Phi(s))|ds$

$\le m\int_0^t g(r(s))ds = r(t) = \sum_{k=1}^\infty \frac{t^k}{k!}\partial_t^k r(0).$

For $\forall |t| < \frac{a}{2b}m$, the majorant taylor series absolutely converge. In order to prove $\Phi(t)$ is also analytic in $t \in [0, \tau)$, where $\tau := \frac{a}{2b}m = \frac{a}{2b}\frac{hq}{hq-q+1}$. We must to prove

$$|\partial_t^k \Phi(0)| \le \partial_t^k r(0), k \ge 1.$$

Supposing that's right, then the taylor series for $\Phi(t)$ has the majorant series. According to Weierstrass M-test, consequently, it converges. First we prove the bound above by computing the following formulas for $k = 1, 2, 3$,

$\partial_t \Phi(t) = mN(\Phi(t)),$

$\partial_t^2 \Phi(t) = mN'(\Phi(t))N(\Phi(t)),$

$\partial_t^3 \Phi(t) =$

$mN''(\Phi(t))N(\Phi(t))N(\Phi(t)) + mN'(\Phi(t))N'(\Phi(t))N(\Phi(t)).$

As a result,

$|\partial_t \Phi(0)| \le m|N(\Phi(0))| \le mg(r(0)) = \partial_t r(0),$

$|\partial_t^2 \Phi(0)| \le m|N'(\Phi(0))||N(\Phi(0))|$

$\le mg'(r(0))g(r(0)) = \partial_t^2 r(0),$

$|\partial_t^3 \Phi(0)|$

$\le m|N''(\Phi(0))||N(\Phi(0))||N(\Phi(0))|$

$+m|N'(\Phi(0))||N'(\Phi(0))||N(\Phi(0))|$

$\le mg''(r(0))g(r(0))g(r(0)) + mg'(r(0))g'(r(0))g(r(0))$

$= \partial_t^3 r(0),$

Generally speaking, for $\forall k \ge 0$,

$$\Phi^{(k+1)}(t) = mP_k(N(\Phi(t))),$$

where $P_k N(\Phi(t))$ is a polynomial of N, and its Fréchet derivatives up to the kth order with positive coefficients. In consequence, we easily get

$|\partial_t^{(k+1)} \Phi(0)| \le m|P_k(N(\Phi(0)))| \le mP_k(|N(\Phi(0))|)$

$$\le mP_k(g(r(0))) \le \partial_t^{(k+1)} r(0),$$

So far, this bound have already concluded the proof.

Corollary 1. For the formula

$$\tau(h) = \frac{a}{2b}\frac{hq}{hq - q + 1}, \tag{3.4.0}$$

the closer the value of h is to zero, the larger the convergence region $\tau(h)$.

proof. Solving

$$\tau(h) > 0,$$

we can obtain

$$h \in (-\infty, 1-\frac{1}{q}) \cup (0, +\infty)$$

The formula (3.4.0) equivalent to

$$\tau(h) = \frac{a}{2b}\frac{1}{1-\frac{1}{h}(1-\frac{1}{q})}, \tag{3.4.1}$$

From (3.4.1), we can get the following conclusions for the convergence region τ.
(i)the closer the value of h is to $+\infty$, the closer to $\frac{a}{2b}$.
(ii)the closer the value of h is to $1-\frac{1}{q}$, the larger the convergence region τ. Eespecially, when $q \to 1^{-}$,$h \to 0^{-}$,$\tau \to +\infty$.

The proof has been finished in (ii) and the conclusion can be showed in FIG.1.

IV. HAM FOR THE NLS EQUATION

Some nonlinear equations often presented as

$$\begin{cases} L(u) = N(u), & t > 0, \\ u(0) = f, \end{cases} \tag{4.1}$$

For example, consider the following continuous NLS equation

$$iu_t = -\frac{1}{2}u_{xx} - |u|^2 u, t > 0, \tag{4.2}$$

where $u(x,t)$ is an amplitude function with the property of $|u|^2 = u\bar{u}$, this equation usually called NLS equation with cubic nonlinearity. Obviously, the equation fits to the above abstract formulation (4.1) with

$$N(u) = \frac{1}{2}i\partial_x^2 u + i|u|^2 u.$$

Do the following transformation: $u(x,t) = F(x)e^{it}$, then use $u(t)$ instead of $F(x)$. what's more, take the initial data $u(0) = 1, \dot{u}(0) = 0$, we will obtain

$$\begin{cases} \ddot{u}(t) = u - 2u^3, & t > 0, \\ \dot{u}(0) = 0, \\ u(0) = 1, \end{cases} \tag{4.3}$$

Eq.(4.1)'s exact solution is $sech(t)$. It is straightforward represent

$$u(t) = \sum_{k=0}^{\infty} a_k t^k, \tag{4.4}$$

under the set of base functions

$$\{t^k, k = 0, 2, 4, ...\},$$

With the aid of the Eq. (4.3) and under the rule of solution expression,we choose the initial approximation

$$u_0(t) = 1,$$

and the auxiliary linear operator

$$L = \frac{\partial^2}{\partial t^2},$$

and

$$\mathcal{N}(\Phi(t,q)) = \frac{\partial^2}{\partial t^2}\Phi(t,q) - \frac{\partial}{\partial t}\Phi(t,q) + 2\Phi^3(t,q),$$

According to the above conditions, we obtain the zero-order deformation equation

$$\begin{cases} (1-q)L[\Phi(t,q) - u_0(t)] = hqH(t)\mathcal{N}[\Phi(t,q)]. \\ \Phi(0,q) = 1, \end{cases} \tag{4.5}$$

From the above Eq. (4.5), we know $\Phi(t,0) = u_0(t)$ and $\Phi(t,1) = u(t)$ when $q = 0$ and $q = 1$. With respect to the embedding parameter q, we define

$$u_k(t) = \frac{1}{k!}\frac{\partial^k}{\partial q^k}\Phi(t,q),$$

and $\Phi(t,q)$ can be expanded in Taylor series

$$\Phi(t,q) = \Phi(0,q) + \sum_{k=0}^{\infty} u_k(t)q^k,$$

Define the following vector

$$\vec{u}_n(t) = \{u_0(t), u_1(t), ..., u_n(t)\}.$$

Then, differentiating the Eq. (4.5) k times with respect to q and dividing by k! at last, we obtain the kth-order deformation equation

$$L[u_k(t) - \chi_k u_{k-1}(t)] = hH(t)R_k[\vec{u}_{k-1}(t)],$$

suffer from the initial condition $u_k(0) = 1$, where

$$R_k[\vec{u}_{k-1}(t)] = \ddot{u}_{k-1}(t) - u_{k-1}(t)$$

$$+2\sum_{i=0}^{k-1}\sum_{j=0}^{k-1-i} u_i(t)u_j(t)u_{k-1-i-j}(t) - (1-\chi_k)$$

and

$$\chi_k = \begin{cases} 0, & k \leq 1, \\ 1, & k > 1. \end{cases}$$

According to the both of the rule of solution expression and the coefficient ergodicity, the corresponding auxiliary function should be determined uniquely

$$H(t) = 1.$$

Then we successively have

$u_1(t) = \frac{1}{2}ht^2,$

$u_2(t) = \frac{1}{2}ht^2 + \frac{1}{2}h^2t^2 + \frac{5}{24}h^2t^4,$

$u_3(t) = \frac{1}{2}ht^2 + h^2t^2 + \frac{1}{2}h^3t^2 + \frac{5}{12}h^2t^4 + \frac{5}{12}h^3t^4 + \frac{61}{720}h^3t^6,$

$\ldots\ldots$

According to the formula $u(t) = \sum_{k=0}^{\infty} u_k(t)$, $u(t)$ could be obtained. As is showed in Figure 2.

V. The Blocked homotopy analysis method

We introduce a new modified method of HAM called blocked homotopy analysis method. If only considered the graph of $h = -1$ from the right graph of Figure 2, we find the approximate solution are very closed to the exact solution in $[0, 1]$ but away from in $[1, 4]$. In generally, compare the approximate with the exact solutions for different value of h, we find that there exists a t_0, such that they could be overlap for any $t \in [0, t_0]$, but be away for $t \in [t_0, \infty]$. Now, let's talk about the idear: First, we can get the approximate solution of one nonlinear equation under the HAM, then we can obtain the graphs of the approximate and the exact solution.Second, we can choose the point t_0 on the overlap section, we know the approximate solution are equal to the exact for any $t \in [0, t_0]$, so we keep the part $t \in [0, t_0]$. Third, we choose t_0 as the starting point and use HAM to the nonlinear equation again to get a new approximate solution and its graph. We can find a point $t_1 (> t_0)$ on the overlap section of the new approximate and exact solution's graphs, then keep the part $t \in [t_0, t_1]$ for the same reason. $t_2, t_3, t_4, ...$ would be find after repeat the above step again and again, where $\{t_n, n = 0, 1, 2, ...\}$ is not unique. At last, the exact solution of the equation would be gotten.

Consider the example of Riticci equation,

$$\dot{u}(t) = 1 - u^2(t), t > 0, \tag{5.1}$$

with initial conditions $u(0) = 0$.

For simplicity, we let $h = -1$ and just take $u(t) \approx t - \frac{1}{3}t^3$ after using the HAM to Eq(5.1) with $u_0(t) = t$, because the picture of the approximate solution $t+t^3$ and the exact solution $tanh(t)$ has already overlap in $[0, 0.5]$, that is to say, [0,0.5] is a valid interval for Riticci equation. As is showed on the left graph of Figure 3.

In generally, use the HAM, we can get $u(t)'s$ approximate solution

$$u(t) \approx (u_0^5 - \tfrac{5}{3}u_0^3 + \tfrac{2}{3}u_0)t^4 + (-u_0^4 + \tfrac{4}{3}u_0^2 - \tfrac{1}{3})t^3$$
$$+(u_0^3 - u_0)t^2 + (1 - u_0^2)t + u_0 := app(t), \tag{5.2}$$

with $h = -1$ around $t = 0$, where $u_0(t)$ is an undetermined initial data. It can be easily find that $u(t) \approx t - \frac{1}{3}t^3$ when $u_0(t) = 0$. Then we use the blocked homotopy analysis method, the figure could be obtained in $t \in [0, \infty]$ with the values of $t_0 = 0.5, t_1 = 1.0, ...$ as showed on the right graph of Figure 3.

Error analysis: Liao tells us that the solution of HAM is analytic and the approximate solution $app(t)'s$ accurate is $O(0.5^4)$ when $u_0(t) \neq 0$, or $O(0.5^3)$ when $u_0(t) = 0$. The first-order derivative of the approximate solution $app(t)'s$ accurate is $O(0.5^3)$ when $u_0(t) \neq 0$, or $O(0.5^2)$ when $u_0(t) = 0$. The second-order derivative of the approximate solution $app(t)'s$ accurate is $O(0.5^2)$ when $u_0(t) \neq 0$, or $O(0.5^1)$ when $u_0(t) = 0$. All of those is showed on Figure 4, and the graph of $app(t), app'(t), app''(t), t \in [0.5, 1.0]$ is showed on Figure 5.

Figure 4 shows that the error of app(t) is satified with the analysis. For instance, when $t = 0.5-$, the error between $app(t)$ and $u(t)$ is $0.003784 (\leq 0.5^3)$, the error between $app(t)$ and $u(t)$ is $0.036448 (\leq 0.5^2)$, the error between $app(t)$ and $u(t)$ is $0.273138 (\leq 0.5^1)$; when $t = 1.0-$, the error between $app(t)$ and $u(t)$ is $0.0006476 (\leq 0.5^4)$, the error between $app(t)$ and $u(t)$ is $0.017533 (\leq 0.5^3)$, the error between $app(t)$ and $u(t)$ is $0.119422 (\leq 0.5^2)$, and so do the others value of t. Moreover, the errors of $app^{(n)}(t+), n = 0, 1, 2$ is smaller than the errors of $app^{(n)}(t-), n = 0, 1, 2$.

Figure 5 shows the errors in $[0.5, 1.0]$. From the graph, we find that the larger value of t, the bigger error of every $app(t), app'(t), app''(t)$. And of course, to any $t \in [0.5, 1.0]$, the error of $app(t)$ is smaller than the error of $app'(t)$ and the error of $app'(t)$ is smaller than the error of $app''(t)$.

VI. Blocked homotopy analysis method for Duffing equation

Let's consider the following Duffing equation as an example,

$$\ddot{u} + \frac{u^3}{1 + u^2} = 0, t > 0, \tag{6.1}$$

with initial conditions $\dot{u}(0) = 0, u(0) = A$. Take $\tau = \omega t$, where ω is the original frequency, and A is the original amplitude. In order to use the Blocked homotopy analysis method, we should apply the HAM to Eq.(6.1) in the first place. Clearly, the solution of Eq.(6.1) can express

$$u(\tau) = \sum_{k=1}^{\infty} c_k cos(k\tau), \tag{6.2}$$

under the set of base functions

$$\{cos(k\tau), k = 1, 2, 3, ...\},$$

where c_k is undetermined. We can choose the linear operator

$$\mathcal{L}[\Phi(\tau, q)] = \omega_0^2 [\frac{\partial}{\partial \tau^2}\Phi(\tau, q) + \Phi(\tau, q)],$$

where $\mathcal{L}$ is satisfied with

$$\mathcal{L}(C_1 sin\tau + C_2 cos\tau) = 0,$$

C_1, C_2 are two constants.

We define the nonlinear operator

$$\mathcal{N}[\Phi(\tau, q), \Omega(q)] = \Omega^2(q)(\frac{\partial}{\partial \tau^2}\Phi(\tau, q))[1+\Phi^2(\tau, q)+\Phi^3(\tau, q)], \tag{6.3}$$

where $\Phi(\tau, q), \Omega(q)$ is a continuous mapping of $u(t), \omega$. We construct the zero order deformation equation

$$(1 - q)\mathcal{L}[\Phi(\tau, q) - u_0(\tau)] = hqH(\tau)\mathcal{N}[\Phi(\tau, q), \Omega(q)], \tag{6.4}$$

with the initial contions

$$\Phi(0, q) = A, \frac{\partial}{\partial \tau}\Phi(\tau, q)|_{\tau=0} = 0$$

where $q \in [0, 1]$ is an embedded variable, $h, H(\tau) \neq 0$, $\mathcal{L}$ is a linear operator, $u_0(\tau)$ is a initial guess solution of $u(\tau)$. With a view of the Eq. (6.1) and the rule of solution expression, we choose $u_0(\tau) = Acos(\tau)$.

When q from 0 increase to 1, we can see

$$\Phi(\tau, 0) = u_0(\tau), \Omega(0) = \omega_0$$

$$\Phi(\tau,1)=u(\tau),\Omega(1)=\omega$$

In other word, when q from 0 increase to 1, $\Phi(\tau,q),\Omega(q)$ change from the initial guess solutions $u_0(\tau),\omega_0$ to the exact solutions $u(\tau),\omega$. Apply the Taylor series expansion to $\Phi(\tau,q)$ and $\Omega(q)$,

$$\Phi(\tau,q)=u_0(\tau)+\sum_{m=1}^{\infty}u_m(\tau)q^m,\Omega(q)=\omega_0+\sum_{m=1}^{\infty}\omega_m q^m$$

where

$$u_m(\tau)=\frac{1}{m!}\frac{\partial^m}{\partial q^m}\Phi(\tau,q)|_{q=0},\omega_m=\frac{1}{m!}\frac{\partial^m}{\partial q^m}\Omega(q)|_{q=0}$$

When $q=1$, we obtain

$$u(\tau)=u_0(\tau)+\sum_{m=1}^{\infty}u_m(\tau),\omega=\omega_0+\sum_{m=1}^{\infty}\omega_m$$

Differentiating the zero order deformation equation m times with respect to q and dividing by $m!$ at last, we obtain the kth-order deformation equation

$$\mathcal{L}[u_m(\tau)-\chi_m u_{m-1}(t)]=hH(t)R_m[u_{m-1},\omega_{m-1}],$$

where

$$R_m[u_{m-1},\omega_{m-1}]=\frac{1}{m!}\frac{\partial^{m-1}}{\partial q^{m-1}}\mathcal{N}(\Phi(t,q),\Omega(q))|_{q=0}, \quad (6.5)$$

and

$$\chi_k=\begin{cases}0, & k\leq 1,\\ 1, & k>1.\end{cases}$$

Put Eq.(6.3) into the above Eq.(6.5), we obtain

$$R_m[u_{m-1},\omega_{m-1}]=\sum_{n=0}^{\varphi(m)}b_n(\omega_{m-1})cos[(2n+1)\tau]$$

where $b_n(\omega_{m-1})$ is based on ω_{m-1} and $\varphi(m)$ is based on m. Then we can get the following formulation

$$R_1=\omega_0^2\ddot{u}_0+\omega_0^2\ddot{u}_0u_0^2+u_0^3$$

$$R_1=b_1cos(\tau)+b_2cos(3\tau)$$

Put $u_0(\tau)=Acos(\tau)$ into the above equations, we get

$$b_1=-\omega_0^2+\frac{3}{4}A^2-\frac{3}{4}A^2\omega_0^2,b_2=\frac{1}{4}(A^3-A^3\omega_0^2)$$

For simplicity we make $H(\tau)=1,b_1=0$, we have $\omega_0=\sqrt{\frac{3A^2}{3A^2+4}}$. At last, we get the M-th approximate solution

$$u(\tau)\approx\sum_{m=0}^{M}u_m(\tau),\omega\approx\sum_{m=0}^{M}\omega_m$$

According to the above HAM, we can use the blocked homotopy analysis method to solve Eq.(6.1).

Figure 5 shows the comparation of Eq.(6.1)'s approximate solution and exact solution, and the comparation of the two solution's first-order derivative on [0,100] when A=0.5.

VII. Conclusion

We have give the proof of the convergence for the powerful analytical method Homotopy analysis method, in the spirit of the Cauchy-Kowalevskaya theorem. We also analysis the parameter h in the contribution of the controlling the convergence region, the results shows that if we adjust the value of h, the convergence regime will changed, as guessed by the prof. Liao who proposed the HAM.

REFERENCES

[1] SJ. Liao, The proposed homotopy analysis technique for the solution of nonlinear problems, Ph.D. Thesis, Shanghai Jiao Tong University, 1992.

[2] SJ. Liao, Int. J. Nonlinear Mech. 30 (1995) 371.

[3] SJ. Liao, Int. J. Nonlinear Mech. 32(5) (1997) 815.

[4] SJ. Liao, Int. J. Non-Linear Mech. 34(4) (1999) 759.

[5] SJ. Liao, Beyond perturbation: introduction to the Homotopy Analysis Method. CRC Press, Boca Raton: Chapman Hall; 2003.

[6] SJ. Liao, Appl. Math. Comput. 147 (2004) 499.

[7] SJ. Liao, Appl. Math. Comput.169 (2005) 1186.

[8] M. Sajid, T. Hayat, and S. Asghar, Comparison between the HAM and HPM solutions of tin film flows of non-Newtonian fluids on a moving belt, Nonlinear Dynamics, online

[9] Allan, F.M., Appl Math Comp. doi:10.1016/j.amc.2006.12.074.

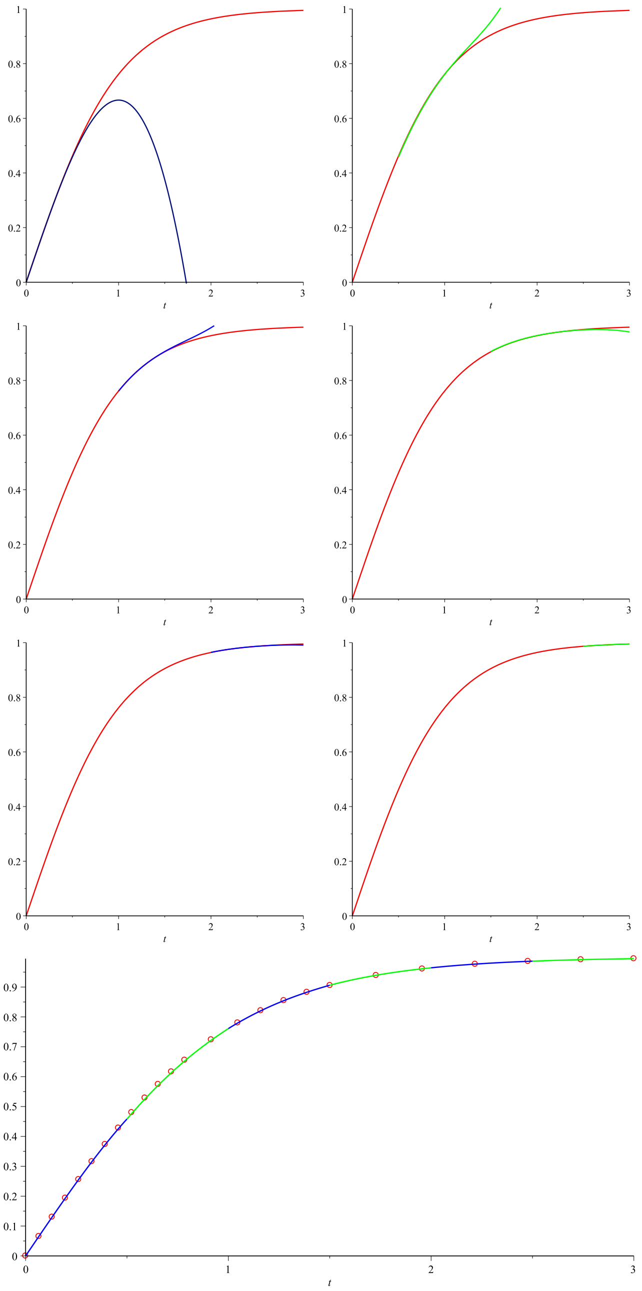
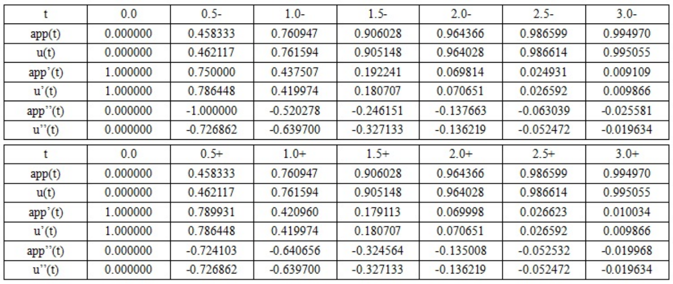


Fig. 3. Left:the blue line is $u(t)'s$ approximate solution $t-\frac{1}{3}t^3$,the red line is $tanh(t)$. Middle:the green line is $u(t)'s$ approximate solution with $u_0 = t_0 - \frac{1}{3}t_0^3, t = t - t_0$ in Eq.(5.2) on (t_0, ∞), where $t_0 = 0.5$. the red line is $tanh(t)$. Right:the point of circle is the exact solution, the blue and green line is the approximate solution on divided sections with $t_0 = 0.5, t_1 = 1.0, t_2 = 1.5, t_3 = 2.0, t_4 = 2.5, t_5 = 3.0$.



t	0.0	0.5-	1.0-	1.5-	2.0-	2.5-	3.0-
app(t)	0.000000	0.458333	0.760947	0.906028	0.964366	0.986599	0.994970
u(t)	0.000000	0.462117	0.761594	0.905148	0.964028	0.986614	0.995055
app'(t)	1.000000	0.750000	0.437507	0.192241	0.069814	0.024931	0.009109
u'(t)	1.000000	0.786448	0.419974	0.180707	0.070651	0.026592	0.009866
app''(t)	0.000000	-1.000000	-0.520278	-0.246151	-0.137663	-0.063039	-0.025581
u''(t)	0.000000	-0.726862	-0.639700	-0.327133	-0.136219	-0.052472	-0.019634

t	0.0	0.5+	1.0+	1.5+	2.0+	2.5+	3.0+
app(t)	0.000000	0.458333	0.760947	0.906028	0.964366	0.986599	0.994970
u(t)	0.000000	0.462117	0.761594	0.905148	0.964028	0.986614	0.995055
app'(t)	1.000000	0.789931	0.420960	0.179113	0.069998	0.026623	0.010034
u'(t)	1.000000	0.786448	0.419974	0.180707	0.070651	0.026592	0.009866
app''(t)	0.000000	-0.724103	-0.640656	-0.324564	-0.135008	-0.052532	-0.019968
u''(t)	0.000000	-0.726862	-0.639700	-0.327133	-0.136219	-0.052472	-0.019634

Fig. 4. $u(t_0+)$ means the right side derivative on $t = t_0$, $u(t_0-)$ means the left side derivative on $t = t_0$. All of the above values is accurate to six decimal places.

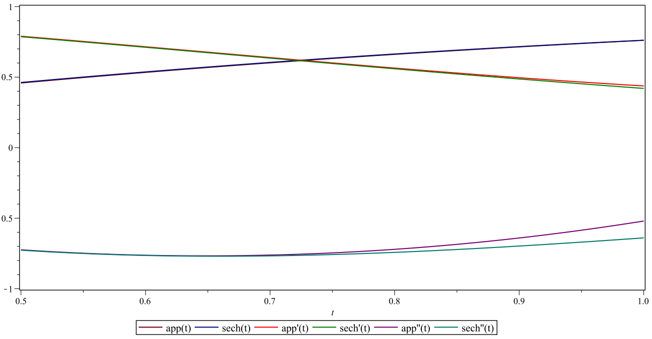


Fig. 5. From top to bottom, the middle two lines is the exact solution and the approximate solution, whose precision is 0.5^4; the first two lines is the first-order derivative of the exact solution and the approximate solution, whose precision is 0.5^3; the last two lines is the second-order derivative of the exact solution and the approximate solution, whose precision is 0.5^2.

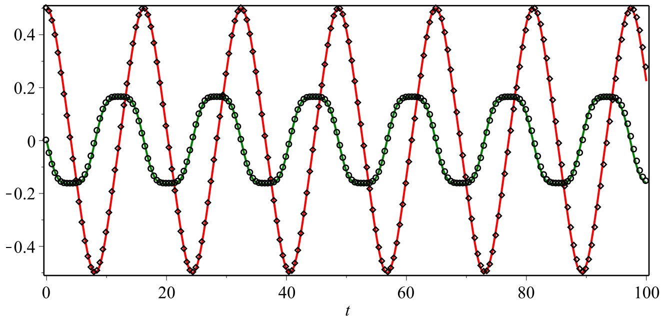


Fig. 6. The rhombuses is Eq.(6.1)'s approximate solution, the red line is the exact solution, the cycles is the first-order derivative of the approximate solution, the green line is the first-order derivative of the exact solution.

A numerical method for solving linear differential equations via Walsh functions

György Gát, and Rodolfo Toledo

***Abstract*—This paper presents a numerical method for solving first order linear ordinary differential equations whose coefficients are not necessarily constant. The method is based on the discretization of the equivalent integral equation with the use of Walsh-Fourier series to obtain a linear system. The solution is approached with a Walsh polynomial whose coefficients are the solution of the linear system. We also deal with the estimation of errors for equations with constant coefficients. Finally, we give an example to illustrate the obtained results.**

***Index Terms*—dyadic convolution matrix, dyadic modulus of continuity, first order linear differential equations, Walsh-Fourier series**

I. Introduction

TE basic idea of using Fourier series for solving differential equations is to assume that the unknown solution can be approximated by the linear combination of basis functions and then this series is substituted into the equation. The aim is to develop a method to choose the series coefficients such that the residuum is minimized. However, the substitution is only possible if the basis functions are derivable and it is not true in case of locally constant orthonormal systems like systems formed by Walsh functions. In this case we substitute the series into the equivalent integral equation with which we avoid the differentiation of basis functions.

In this paper we deal with a numerical solution of the Cauchy problem

$$\begin{gathered} y' + p(x)y = q(x) \\ y(0) = \eta \end{gathered}$$

where $p, q\colon [0,1[\to \mathbf{R}$ is a continuous and integrable function. This problem is equivalent to the following integral equation:

$$y(x) = \eta + \int_0^x q(t) - p(t)y(t)\,dt \qquad (0 \leq x < 1).$$

We propose to approach the solution of integral equation above by the Walsh polynomials

$$\overline{y}_n(x) = \sum_{k=0}^{2^n-1} c_k \omega_k(x),$$

satisfying the equation

$$\overline{y}_n(x) = \eta + S_{2^n} \int_0^x S_{2^n} q(t) - S_{2^n} p(t) \overline{y}_n(t)\,dt$$

where $0 \leq x < 1$, ω_k is the kth Walsh-Paley function and $S_{2^n} f$ are the 2^n-partial sums of Walsh-Fourier series of an integrable function f on the interval $[0,1[$.

In the next three sections we introduce some basic notations and statements that we use throughout this paper.

II. The Walsh-Paley system

Every $n \in \mathbf{N}$ can be uniquely expressed as

$$n = \sum_{k=0}^{\infty} n_k 2^k,$$

where $n_k = 0$ or $n_k = 1$ for all $k \in \mathbf{N}$. This allows us to say that the sequence $(n_0, n_1, \dots)$ is the dyadic expansion of n. Similarly, the dyadic expansion $(x_0, x_1, \dots)$ of a real number $x \in [0,1[$ is given by the sum

$$x = \sum_{k=0}^{\infty} \frac{x_k}{2^{k+1}},$$

where $x_k = 0$ or $x_k = 1$ for all $k \in \mathbf{N}$. This expansion is not unique if x is a dyadic rational, i.e. x is a number of the form $\frac{i}{2^k}$, where $i, k \in \mathbf{N}$ and $0 \leq i < 2^k$. When this situation occurs we choose the expansion terminates in zeros.

The Walsh-Paley function ω_n is obtained by the finite product of Rademacher functions

$$r_k(x) := (-1)^{x_k} \qquad (x \in [0,1[,\ k \in \mathbf{N}),$$

namely

$$\omega_n(x) := \prod_{k=0}^{\infty} r_k^{n_k}(x) \qquad (x \in [0,1[, n \in \mathbf{N}).$$

The Walsh-Paley system is an orthonormal and complete system on $L^1([0,1[)$.

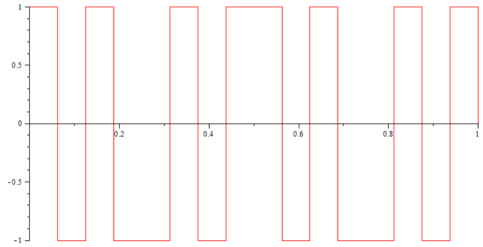


Fig. 1. The Walsh-Paley function ω_{10}

Research supported by project TÁMOP-4.2.2.A-11/1/KONV-2012-0051.

György Gát (email: gatgy@nyf.hu) and Rodolfo Toledo (email: toledo@nyf.hu) are with the Institute of Mathematics and Computer Science, College of Nyíregyháza, P.O.Box 166., H–4400, Hungary.

For an integrable function f on the interval $[0,1[$ we define the Fourier coefficients and partial sums of Fourier series by

$$\widehat{f}_k := \int_0^1 f(x)\omega_k(x)\,dx \quad (k \in \mathbf{N}),$$

$$S_n f(x) := \sum_{k=0}^{n-1} \widehat{f}_k \omega_k(x) \quad (n \in \mathbf{N}, x \in [0,1[).$$

It is important to note that the partial sums $S_{2^n} f$ can be written as

$$S_{2^n} f(x) = 2^n \int_{I_n(x)} f(y)\,dy$$

where the sets

$$I_k(i) := \left[\frac{i-1}{2^k}, \frac{i}{2^k}\right[\qquad (i = 1, \dots, 2^k)$$

are called dyadic intervals and $I_n(x)$ is the dyadic interval which contains the value of x. For this reason, the operator S_{2^n} is the conditional expectation with respect to the σ-algebra generated by the sets $I_n(x)$, $x \in [0,1[$. Thus, by the martingale convergence theorem we obtain that $S_{2^n} f$ converge to f in L^p-norm and a.e. for all functions $f \in L^p([0,1[)$, $p \geq 1$ (see [4] p. 29).

III. The dyadic modulus of continuity

The topology generated by the collection of dyadic intervals on $[0,1[$ is called dyadic topology. Define the dyadic sum of two numbers $x, y \in [0,1[$ with expansion $(x_0, x_1, \dots)$ and $(y_0, y_1, \dots)$ respectively by

$$x \dotplus y := \sum_{k=0}^{\infty} |x_k - y_k| 2^{-(k+1)}.$$

Then $\rho(x,y) := x \dotplus y$ is a metric on $[0,1[$ which generates the dyadic topology.

Let C_W be the set of real-valued functions on the interval $[0,1[$ which are continuous at every dyadic irrational, continuous from the right on $[0,1[$ and have a finite limit from the left on $]0,1]$, all this on the usual topology. It is not hard to see that every continuous function $f \colon [0,1[\to \mathbf{R}$ on the usual topology belongs to C_W if f has a finite limit from the left of 1. Moreover, every function in C_W is continuous on the dyadic topology. On the other hand, every continuous function $f \colon [0,1[\to \mathbf{R}$ on the usual topology is also continuous on the dyadic topology (see [4] p. 11), but it is necessary to have a finite limit from the left of 1 in order to be in C_W.

A function belongs to the Walsh-Paley system is called Walsh function. The finite linear combinations of Walsh functions

$$f(x) = \sum_{k=0}^{n} a_k \omega_k(x)$$

are called Walsh polynomials. Every Walsh polynomial is a dyadic step function and vice versa. Every continuous function on the interval $[0,1[$ can be approximated by Walsh polynomials at every point, but this can be done uniformly only for functions belongs to C_W.

Define the dyadic modulus of continuity of an $f \in C_W$ by

$$\omega_n f := \sup\{|f(x \dotplus h) - f(x)| \colon x \in [0,1[, 0 \leq h < 2^{-n}\}$$

and the local modulus of continuity on the dyadic interval $I_n(i)$ of a continuous function f on the dyadic topology by

$$\omega_{n,i} f := \sup\{|f(x \dotplus h) - f(x)| \colon x \in I_n(i), 0 \leq h < 2^{-n}\}.$$

for all $i = 1, 2, \dots, 2^n$. Not every $f \colon [0,1[\to \mathbf{R}$ continuous function on the usual topology has a finite modulus of continuity, it is also necessary the existence of the limit of f from the left of 1.

The dyadic modulus of continuity can be written as follows

$$\omega_n f := \sup\{|f(x_1) - f(x_2)| \colon x_1, x_2 \in I_n(i), i = 1, 2, \dots, 2^n\}.$$

and similarly,

$$\omega_{n,i} f := \sup\{|f(x_1) - f(x_2)| \colon x_1, x_2 \in I_n(i)\}.$$

for all $i = 1, 2, \dots, 2^n$. Obviously,

$$\omega_n f = \max\{\omega_{n,i} f \colon i = 1, 2, \dots, 2^n\}.$$

Since every function $f \in C_W$ is uniformly continuous on the interval $[0,1[$ with respect to the dyadic topology, we have $\omega_n f \searrow 0$. If we have the sequence of dyadic intervals $I_k(i_k) \supseteq I_{k+1}(i_{k+1}) \supseteq \dots$ and f is a continuous function on $[0,1[$ such that $\omega_{k,i_k} f$ is finite then $\omega_{n,i_n} f \searrow 0$.

For every $f \in C_W$ and $x \in [0,1[$ we have (see [4])

$$\frac{1}{2}\omega_n f \leq |f(x) - S_{2^n} f(x)| \leq \omega_n f. \tag{1}$$

Similarly, if $x \in I_n(i)$ we have

$$\frac{1}{2}\omega_{n,i} f \leq |f(x) - S_{2^n} f(x)| \leq \omega_{n,i} f. \tag{2}$$

Denote $K_f := \sup\{|f(x)| \colon x \in [0,1[\}$. If the function f satisfies the Lipschitz condition, i.e. there is a constant L such that

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$$

for all $x_1, x_2 \in [0,1[$, then $\omega_n f \leq \frac{L}{2^n}$. Moreover, if f is a derivable function on $[0,1[$ with bounded derivative, then $L = K_{f'}$ and $\omega_n f \leq \frac{K_{f'}}{2^n}$. Similar statements are also true for the local moduli of continuity.

IV. The triangular functions

Denote the triangular functions by

$$J_k(x) := \int_0^x \omega_k(t)\,dt \qquad (k \in \mathbf{N}, 0 \leq x < 1).$$

and the Walsh series of these functions by

$$J_k(x) = \sum_{j=0}^{\infty} \widehat{J}_{k,j} \omega_j(x).$$

Coefficients $\widehat{J}_{k,j}$ often take the value 0, in particular they have the following properties:

$$\begin{aligned}
&\widehat{J}_{0,0} = \frac{1}{2}, \\
&\widehat{J}_{k,j} = \frac{1}{2^{n+1}} && \text{if } 0 \leq j < 2^{n-1} \leq k < 2^n,\ k = 2^{n-1} + j, \\
&\widehat{J}_{k,j} = 0 && \text{if } 0 \leq j < 2^{n-1} \leq k < 2^n,\ k \neq 2^{n-1} + j, \\
&\widehat{J}_{k,j} = -\frac{1}{2^{n+1}} && \text{if } 0 \leq k < 2^{n-1} \leq j < 2^n,\ k = j - 2^{n-1}, \\
&\widehat{J}_{k,j} = 0 && \text{if } 0 \leq k < 2^{n-1} \leq j < 2^n,\ k \neq j - 2^{n-1}, \\
&\widehat{J}_{k,j} = 0 && \text{if } 2^{n-1} \leq k, j < 2^n,
\end{aligned}$$

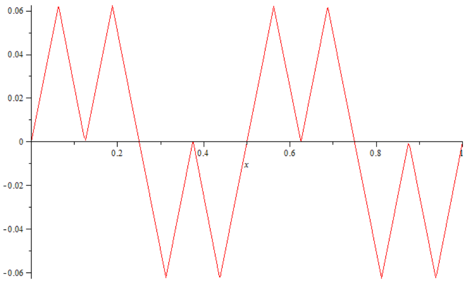


Fig. 2. The triangular function J_{10}

for every $n \in \mathbf{N}$. We can obtain the values above directly by the Fine's formulae (see [3])

$$J_0(x) = \frac{1}{2} - \frac{1}{4}\sum_{i=0}^{\infty} \frac{1}{2^i}\omega_{2^i}(x)$$

and

$$J_k(x) = \frac{1}{2^{n+1}}\left(\omega_l(x) - \sum_{i=1}^{\infty}\frac{1}{2^i}\omega_{2^{n-1+i}+k}(x)\right) \qquad (3)$$

for all $k = 2^{n-1} + l$, $0 \le l < 2^{n-1}$. Indeed, if $2^{n-1} \le k < 2^n$ then $2^{n-1+i} + k \ge 2^n$ for all $i \ge 1$, so in (3) only the coefficient with index $l = 2^{n-1} - k$ is not zero if we only consider indexes less than 2^n. Hence, $\widehat{J}_{k,j} \neq 0$ only if $j = l$, that is $k = 2^{n-1} + j$ and $\widehat{J}_{k,j} = \frac{1}{2^{n+1}}$. On the other hand, if $k < 2^{n-1}$ and $k = 2^{n'} + l$, where $0 \le l < 2^{n'}$, then $2^{n-1} \le 2^{n'+i} + k < 2^n$ only for $i = n - 1 - n'$. For this reason, the Walsh series

$$J_k(x) = \frac{1}{2^{n'+1}}\left(\omega_l(x) - \sum_{i=1}^{\infty}\frac{1}{2^i}\omega_{2^{n'+i}+k}(x)\right).$$

only have one non-zero coefficient with index j such that $2^{n-1} \le j < 2^n$. In this case, $j = 2^{n'+(n-1-n')} + k = 2^{n-1} + k$ and $\widehat{J}_{k,j} = \frac{1}{2^{n'+1}} \cdot \frac{-1}{2^{n-1-n'}} = -\frac{1}{2^{n+1}}$.

Denote by $\widehat{J}^{(n)}$ the matrix of size 2^n with entries $\widehat{J}_{k,j}$, where $k, j = 0, 1, \ldots, 2^n - 1$. The properties of $\widehat{J}_{k,j}$ allow us to construct the matrices $\widehat{J}^{(n)}$ in an easier way as follows:

$$\left(\begin{array}{cc|c|c} \frac{1}{2} & \ddots & -\frac{1}{2^n}\mathcal{I}_{2^{n-2}} & \\ \ddots & \ddots & & -\frac{1}{2^{n+1}}\mathcal{I}_{2^{n-1}} \\ \hline \multicolumn{2}{c|}{\frac{1}{2^n}\mathcal{I}_{2^{n-2}}} & 0_{2^{n-2}} & \\ \hline \multicolumn{3}{c|}{\frac{1}{2^{n+1}}\mathcal{I}_{2^{n-1}}} & 0_{2^{n-1}} \end{array}\right)$$

where $\mathcal{I}_j$ and 0_j are the identity and null matrix of size j.

For example

$$\widehat{J}^{(3)} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{8} & 0 & -\frac{1}{16} & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & -\frac{1}{8} & 0 & -\frac{1}{16} & 0 & 0 \\ \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{16} & 0 \\ 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{16} \\ \frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{16} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{16} & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Note that the matrices $\widehat{J}^{(n)}$ can be constructed by the iteration

$$\widehat{J}^{(0)} = \left(\frac{1}{2}\right), \qquad \widehat{J}^{(n)} = \left(\begin{array}{c|c} \widehat{J}^{(n-1)} & -\frac{1}{2^{n+1}}\mathcal{I}_{2^{n-1}} \\ \hline \frac{1}{2^{n+1}}\mathcal{I}_{2^{n-1}} & 0_{2^{n-1}} \end{array}\right).$$

V. The numerical solution of the integral equation

Our aims is to approach the solution of the integral equation

$$y(x) = \eta + \int_0^x q(t) - p(t)y(t)\,dt \qquad (0 \le x < 1) \qquad (4)$$

with the Walsh polynomials

$$\overline{y}_n(x) = \sum_{k=0}^{2^n-1} c_k\omega_k(x) \qquad (5)$$

which satisfy the equation

$$\overline{y}_n(x) = \eta + S_{2^n}\int_0^x S_{2^n}q(t) - S_{2^n}p(t)\overline{y}_n(t)\,dt \qquad (6)$$

where $0 \le x < 1$, and n is a positive integer. We suppose that the functions p and q are continuous on the interval $[0, 1[$ and

$$\int_0^1 p(t)\,dt < \infty, \qquad \int_0^1 q(t)\,dt < \infty.$$

We introduce the following column vectors:

$$\begin{aligned} \mathbf{c}^\top &:= (c_0, c_1, \ldots, c_{2^n-1}), \\ \widehat{\mathbf{p}}^\top &:= (\widehat{p}_0, \widehat{p}_1, \ldots, \widehat{p}_{2^n-1}), \\ \widehat{\mathbf{q}}^\top &:= (\widehat{q}_0, \widehat{q}_1, \ldots, \widehat{q}_{2^n-1}), \\ \boldsymbol{\omega}(\mathbf{x})^\top &:= (\omega_0(x), \omega_1(x), \ldots, \omega_{2^n-1}(x)), \\ \mathbf{e_0}^\top &:= (1, 0, \ldots, 0) \text{ with size } 2^n \end{aligned}$$

and the matrix

$$P := (\widehat{p}_{i\oplus j})_{i,j=0}^{2^n-1}.$$

Matrix P is called a dyadic convolution matrix. These symmetric matrices are useful in our computation by the fact

$$\begin{aligned}\boldsymbol{\omega}(\mathbf{t})^\top \widehat{\mathbf{p}} \boldsymbol{\omega}(\mathbf{t})^\top \mathbf{c} &= \sum_{k=0}^{2^n-1} \widehat{p}_k \omega_k(t) \sum_{j=0}^{2^n-1} c_j \omega_j(t)\\ &= \sum_{k,j=0}^{2^n-1} \widehat{p}_k c_j \omega_k(t)\omega_j(t)\\ &= \sum_{k,j=0}^{2^n-1} \widehat{p}_k c_j \omega_{k\oplus j}(t) =\\ &= \sum_{i,j=0}^{2^n-1} \widehat{p}_{i\oplus j} c_j \omega_i(t)\\ &= \boldsymbol{\omega}(\mathbf{t})^\top P^\top \mathbf{c}\end{aligned}$$

from which we can write (6) as the matrix equality

$$\begin{aligned}\boldsymbol{\omega}(\mathbf{x})^\top \mathbf{c} &= \eta \mathbf{e_0} + S_{2^n} \int_0^x \boldsymbol{\omega}(\mathbf{t})^\top \widehat{\mathbf{q}} - \boldsymbol{\omega}(\mathbf{t})^\top \widehat{\mathbf{p}} \boldsymbol{\omega}(\mathbf{t})^\top \mathbf{c}\, dt\\ &= \eta \mathbf{e_0} + S_{2^n} \int_0^x \boldsymbol{\omega}(\mathbf{t})^\top \widehat{\mathbf{q}} - \boldsymbol{\omega}(\mathbf{t})^\top P\mathbf{c}\, dt\\ &= \eta \mathbf{e_0} + S_{2^n} \int_0^x \boldsymbol{\omega}(\mathbf{t})^\top\, dt(\widehat{\mathbf{q}} - P\mathbf{c})\\ &= \boldsymbol{\omega}(\mathbf{x})^\top \eta \mathbf{e_0} + \boldsymbol{\omega}(\mathbf{x})^\top \widehat{J}^{(n)\top}(\widehat{\mathbf{q}} - P\mathbf{c})\\ &= \boldsymbol{\omega}(\mathbf{x})^\top (\eta \mathbf{e_0} + \widehat{J}^{(n)\top}(\widehat{\mathbf{q}} - P\mathbf{c}))\end{aligned} \tag{7}$$

at every point of $[0,1[$. Thus, (7) also holds for the coefficients of the above Walsh polynomials from which we obtain the linear system

$$\mathbf{c} = \eta \mathbf{e_0} + \widehat{J}^{(n)\top}(\widehat{\mathbf{q}} - P\mathbf{c})$$

involving the variables $c_0, c_1, \ldots, c_{2^n-1}$ and it can be written as follows

$$(\mathcal{I}_{2^n} + \widehat{J}^{(n)\top} P)\mathbf{c} = \eta \mathbf{e_0} + \widehat{J}^{(n)\top} \widehat{\mathbf{q}}. \tag{8}$$

The prove of the forthcoming lemma will be publish elsewhere.

Lemma 1. *Let $\widetilde{p}_n := S_{2^n} p$ be the 2^n partial sums of the Walsh series of the function p. Then*

$$\det(\mathcal{I}_{2^n} + \widehat{J}^{(n)\top} P) = \prod_{i=0}^{2^n-1} \left(1 + \frac{\widetilde{p}_n(\frac{i}{2^n})}{2^{n+1}}\right).$$

By Lemma 1 the linear system (8) has an unique solution given by the formula

$$\mathbf{c} = (\mathcal{I}_{2^n} + \widehat{J}^{(n)\top} P)^{-1}(\eta \mathbf{e_0} + \widehat{J}^\top \widehat{\mathbf{q}}) \tag{9}$$

if $2^{n+1} \neq -\widetilde{p}_n(\frac{i}{2^n})$ for all $i = 0, 1, \ldots, 2^n - 1$. Thus, if we suppose that the function p is bounded on the interval $[0,1[$ and

$$K_p := \sup\{|p(x)| \colon x \in [0,1[\}$$

then for every n integer such that

$$2^{n+1} > K_p$$

there exists an unique Walsh polynomial $\overline{y}_n$ which satisfies (6).

VI. Estimation of errors for equations with constant coefficients

In this section we suppose that $p(x) = a$ for all $x \in [0,1[$ where a is a real constant. On the other hand, $q\colon [0,1[\to \mathbf{R}$ has a finite limit from the left of 1, so it can be extended to a continuous function on the interval $[0,1]$. Our aim is to find an upper estimation of the difference $|y(x) - \overline{y}_n(x)|$ at every point $x \in [0,1[$, where y is the unique solution of the integral equation which is given by the formula

$$y(x) = e^{-ax}\left(\eta + \int_0^x q(t)e^{at}\, dt\right) \qquad (0 \le x < 1). \tag{10}$$

Our assumption with respect to the function q ensures that the moduli of continuity $\omega_n q$ and $\omega_n y$ are finite. We will do the estimation in two steps according to

$$|y(x) - \overline{y}_n(x)| \le |y(x) - S_{2^n}(x)| + |S_{2^n}(x) - \overline{y}_n(x)|.$$

Let i be the positive integer for which $x \in I_n(i)$ holds. From the integral equation (4) we have

$$S_{2^n} y(x) = \eta + S_{2^n} \int_0^x q(t) - ay(t)\, dt$$

and by (2)

$$|y(x) - S_{2^n} y(x)| \le \omega_{n,i} y.$$

Let $x_1, x_2 \in I_n(i)$. By the Lagrange's mean value theorem there is a value ζ between x_1 and x_2 such that

$$|y(x_1) - y(x_2)| = |y'(\zeta)||x_1 - x_2| < |y'(\zeta)|\frac{1}{2^n}$$

By (10) we obtain

$$y'(\zeta) = q(\zeta) - ay(\zeta) = q(\zeta) - a\eta e^{-a\zeta} - ae^{-a\zeta}\int_0^\zeta q(t)e^{at}\, dt.$$

Thus,

$$|y'(\zeta)| \le K_{q,i} + |a||\eta|e^{-a\zeta} + |a|e^{-a\zeta} K^*_{q,i} \int_0^\zeta e^{at}\, dt.$$

where $K_{q,i}$ is the supremum of $|q|$ on the interval $I_n(i)$, and $K^*_{q,i}$ the supremum of $|q|$ on the interval $[0, \frac{i}{2^n}]$. If $a = 0$ then $|y'(\zeta)| \le K_{q,i}$. If $a \neq 0$ then

$$\begin{aligned}|y'(\zeta)| &\le K_{q,i} + |a||\eta|e^{-a\zeta} + |a|e^{-a\zeta} K^*_{q,i}\frac{e^{a\zeta} - 1}{a} =\\ &= K_{q,i} + |a||\eta|e^{-a\zeta} + \operatorname{sgn}(a) K^*_{q,i}(1 - e^{-a\zeta}).\end{aligned}$$

Hence we obtain the following estimations

$$\omega_n y \le \begin{cases} \frac{1}{2^n} K_q & \text{if } a = 0,\\ \frac{1}{2^n}(2K_q + a|\eta|) & \text{if } a > 0,\\ \frac{1}{2^n} e^{-a}(K_q - a|\eta|) & \text{if } a < 0,\end{cases} \tag{11}$$

where K_q the supremum of $|q|$ on the whole interval $[0,1[$.

In the second part of the estimation we deal with the absolute value of

$$z_n(x) := \overline{y}_n(x) - S_{2^n} y(x).$$

By the definitions of $\overline{y}_n$ and $S_{2^n}y$ we obtain immediately

$$\begin{aligned} z_n(x) =& S_{2^n}\int_0^x S_{2^n}q(t)-q(t)\,dt + aS_{2^n}\int_0^x y(t)-\overline{y}_n(t)\,dt \\ =& S_{2^n}\int_0^x S_{2^n}q(t)-q(t)\,dt + aS_{2^n}\int_0^x y(t)-S_{2^n}y(t)\,dt \\ & - aS_{2^n}\int_0^x z_n(t)\,dt. \end{aligned}$$

Since z_n is constant on the dyadic intervals $I_n(i)$, we have

$$z_n(x) = \sum_{k=1}^{2^n} z_n(\frac{k-1}{2^n})\chi_{I_n(k)}(x)$$

for all $x \in [0,1[$, where χ_A be the characteristic function of the set A. It is not hard to see that

$$S_{2^n}\int_0^x \chi_{I_n(i)}(t)\,dt = \begin{cases} 0 & \text{if } 0 \le x < \frac{i-1}{2^n}, \\ \frac{1}{2^{n+1}} & \text{if } \frac{i-1}{2^n} \le x < \frac{i}{2^n}, \\ \frac{1}{2^n} & \text{if } \frac{i}{2^n} \le x < 1. \end{cases}$$

Thus, if $I_n(i)$ is the dyadic interval which contains x we obtain

$$S_{2^n}\int_0^x z_n(t) = \frac{1}{2^n}\sum_{k=1}^{i-1} z_n(\frac{k-1}{2^n}) + \frac{1}{2^{n+1}}z_n(\frac{i-1}{2^n})$$

from which we have

$$\begin{aligned} z_n(\frac{i-1}{2^n}) =& S_{2^n}\int_0^x S_{2^n}q(t)-q(t)\,dt \\ & + aS_{2^n}\int_0^x y(t)-S_{2^n}y(t)\,dt \\ & - \frac{a}{2^n}\sum_{k=1}^{i-1} z_n(\frac{k-1}{2^n}) - \frac{a}{2^{n+1}}z_n(\frac{i-1}{2^n}) \end{aligned}$$

and

$$\begin{aligned} \Big(1+\frac{a}{2^{n+1}}\Big)\,z_n(\frac{i-1}{2^n}) =& S_{2^n}\int_0^x S_{2^n}q(t)-q(t)\,dt \\ & + aS_{2^n}\int_0^x y(t)-S_{2^n}y(t)\,dt \\ & - \frac{a}{2^n}\sum_{k=1}^{i-1} z_n(\frac{k-1}{2^n}). \end{aligned}$$

Suppose that $2^{n+1} \ge -3a$. Denote by M_n the supremum of

$$\frac{1}{1+\frac{a}{2^{n+1}}}\left|S_{2^n}\int_0^x S_{2^n}q(t)-q(t)\,dt + aS_{2^n}\int_0^x y(t)-S_{2^n}y(t)\,dt\right|$$

on the interval $[0,1[$, $z_n^* := |z_n|$ and

$$b_n := \frac{\frac{|a|}{2^n}}{1+\frac{a}{2^{n+1}}} = \frac{|a|}{2^n+\frac{a}{2}}$$

With the new notations we obtain

$$z_n^*(\frac{i-1}{2^n}) \le M_n + b_n\sum_{k=1}^{i-1} z_n^*(\frac{k-1}{2^n}),$$

from which by induction we can prove that

$$z_n^*(\frac{i-1}{2^n}) \le M_n(1+b_n)^{i-1} \le M_n(1+b_n)^{2^n}.$$

If $a \ge 0$ then

$$(1+b_n)^{2^n} = \left(1+\frac{a}{2^n+\frac{a}{2}}\right)^{2^n} \le \left(1+\frac{a}{2^n}\right)^{2^n} < e^a.$$

If $a < 0$ then

$$\begin{aligned} (1+b_n)^{2^n} &= \left(1+\frac{|a|}{2^n-\frac{|a|}{2}}\right)^{2^n} \\ &= \left(1+\frac{|a|}{2^n-\frac{|a|}{2}}\right)^{2^n-\frac{|a|}{2}}\left(1+\frac{|a|}{2^n-\frac{|a|}{2}}\right)^{\frac{|a|}{2}} \\ &\le \left(1+\frac{|a|}{2^n-\frac{|a|}{2}}\right)^{2^n-\frac{|a|}{2}}\left(1+\frac{|a|}{|a|}\right)^{\frac{|a|}{2}} \\ &< e^{|a|}2^{\frac{|a|}{2}} = (\sqrt{2}e)^{|a|}, \end{aligned}$$

since $2^n - \frac{|a|}{2} \ge |a|$ according with the assumption $2^{n+1} \ge -3a$.

With respect to the estimation of M_n note that for all $f \in C_W$ and $j = 1,2,\ldots,2^n$ we have

$$\int_0^{\frac{j-1}{2^n}} S_{2^n}f(t)-f(t)\,dt = 0 \qquad (i = 1,2,\ldots,2^n),$$

hence if $s \in I_n(j)$ we have

$$\begin{aligned} \left|S_{2^n}\int_0^s S_{2^n}f(t)-f(t)\,dt\right| &= \left|S_{2^n}\int_{\frac{j-1}{2^n}}^s S_{2^n}f(t)-f(t)\,dt\right| \\ &\le S_{2^n}\int_{\frac{j-1}{2^n}}^{\frac{j}{2^n}} |S_{2^n}f(t)-f(t)|\,dt \\ &\le \frac{1}{2^n}\omega_{n,j}f \le \frac{1}{2^n}\omega_n f. \end{aligned}$$

For this reason

$$\begin{aligned} M_n &\le \frac{1}{1+\frac{a}{2^{n+1}}}\cdot\frac{1}{2^n}\left(\omega_n q + |a|\omega_n y\right) \\ &= \frac{1}{2^n+\frac{a}{2}}\left(\omega_n q + |a|\omega_n y\right) \end{aligned}$$

In summary, if $2^{n+1} \ge -3a$ we have

$$|\overline{y}_n(x) - S_{2^n}y(x)| \le \begin{cases} \frac{1}{2^n+\frac{a}{2}}\left(\omega_n q + |a|\omega_n y\right)e^a & \text{if } a \ge 0, \\ \frac{1}{2^n+\frac{a}{2}}\left(\omega_n q + |a|\omega_n y\right)(\sqrt{2}e)^{-a} & \text{if } a < 0, \end{cases} \tag{12}$$

for all $x \in [0,1[$. Note that the difference between $\overline{y}(x)$ and $S_{2^n}y(x)$ tends very fast to zero, specially if q satisfies the Lipschitz condition, because in this case both $\omega_n q$ and $\omega_n y$ are $O(\frac{1}{2^n})$.

An example

In this section we test the effectiveness of the developed method. Consider the following Cauchy problem with constant coefficients.

$$\begin{aligned} y' + y &= (x+1)^2, \\ y(0) &= 1. \end{aligned} \tag{13}$$

The solution of the Cauchy problem is $y(x) = x^2 + 1$. Figure 3 shows the approximation of $\overline{y}_n$ to the solution y in case of $n = 5$.

Note that the function $q(x) = (x+1)^2$ is continuous on the close interval $[0,1]$, so we can use the estimations (11) and (12), denoting them by E_1 and E_2 respectively. Moreover, in E_2 we estimate the modulus of continuity of q by

$$\omega_n q \le \frac{K_{q'}}{2^n} = \frac{4}{2^n}.$$

Table I contains the errors of the approximation and the estimations E_1 and E_2 for n from 3 to 10. Note that the first error halves and the second error is reduced to quarter with the increment of n.

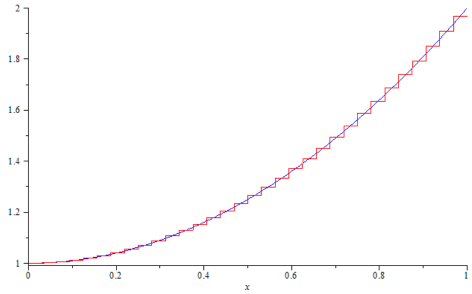


Fig. 3. Approximation of $\overline{y}_5$ to the solution of the Cauchy problem (13)

.

TABLE I
ERRORS RELATED TO THE CAUCHY PROBLEM (13).

n	$\lvert y - \overline{y}_n\rvert$	$\lvert y - S_{2^n}y\rvert$	E_1	$\lvert \overline{y}_n - S_{2^n}y\rvert$	E_2
3	0.1187	0.1197	1.1250	0.0024	0.5196
4	0.0609	0.0611	0.5625	0.0006	0.1338
5	0.0308	0.0309	0.2812	0.00016	0.0339
6	0.015508	0.015523	0.140625	0.000040	0.008560
7	0.007775	0.007772	0.070312	0.000010	0.002148
8	0.00389701	0.00389607	0.03515625	0.00000253	0.00053815
9	0.00195081	0.00195058	0.01757812	0.00000063	0.00013467
10	0.00097598	0.00097592	0.00878906	0.00000015	0.00003368

REFERENCES

[1] C. F. Chen and C. H. Hsiao, *A state-space approach to Walsh series solution of linear systems*, Int. J. Systems Sci, **vol. 6, no. 9** 833-858

[2] C. F. Chen and C. H. Hsiao, *Walsh series analysis in optimal control*, Int. J. Control, **vol. 21, no. 6** (1975), 881-897.

[3] N. J. Fine, *On the Walsh functions*, Trans. Am. Math. Soc. **65** (1949), 372–414.

[4] F. Schipp, W. R. Wade, P. Simon and J. Pál, Walsh series,”An introduction to dyadic harmonic analysis”, Adam Hilger, Bristol and New York, 1990.

User profile based Quality of Experience

Silvia Canale, Francisco Facchinei, Raffaele Gambuti, Laura Palagi and Vincenzo Suraci

***Abstract*—**The Quality of Experience (QoE) is a subjective measure of the quality experienced by an user with respect to a service or a class of services. This measure takes into account a pervasive and holistic evaluation of the service as a whole. It typically differs from objective and structured measure of quality parameters subject to a service provider’s control. In this paper we consider the problem of identifying general behavioural profiles with respect to a generic service starting from raw data describing the user’s feedback in different circumstances. The aim is that of analyzing these profiles in order to recognize what are the most influencing components for QoE.

***Keywords*—**Adaptive evaluation system, Quality of Experience, User profiling.

I. Introduction

THE problem of definition of reliable measures for Quality of Experience (QoE) has been faced in a number of practical applications due to the massive development of resource demanding multimedia services over Internet representing a continuous challenge for service providers. Suitable optimization strategies are made possible by an accurate evaluation of the user’s QoE so to minimize network resource and, at the same time, to satisfy the actual user’s QoE [10]. Differently from user and customer experience, evaluating the quality perceived by the user with respect to a given service is still an open issue from both the theoretical and the practical point of view. Though there is no doubt on the fact that Quality of Service (QoS) parameters like delay, jitter, and packet loss, directly affect the quality perceived by the user, a reliable measurement system to provide a completely QoS based function representing the user’s QoE is still missing. Both subjective (e.g. the Mean Opinion Score) and objective measurement systems have been proposed in the last decade.

In this work we consider a QoE measure system for a telecommunication service, where the QoE measure is usually strongly related to the Quality of Service (QoS) parameters [1]-[4]. In fact, several approaches show the dependency of QoE from QoS parameters according to different services [5]-[8]. Though in these approaches the role of explicit feedback from the user plays an important role, the fact that similar users can provide similar feedbacks about the perceived quality has not been explicitly considered so far. In this paper we cope with the problem of automatically identifying distinct user profiles on the basis of subjective user feedbacks with respect to a generic service. In order to do that, we propose a data model that extracts automatically significant features from raw data time series and a clustering based approach for the identification of the most significant user profiles.

The paper is organized as follows: Section II presents the Future Internet architecture proposed within the FP7 project FI-WARE [13] and the Italian project PLATINO [14]; Section III defines the parameters considered in the QoE evaluation system; Section IV defines the measurement process; Section V draws the conclusions.

II. Future Internet Core Platform Architecture

This section gives a high level overview of the Future Internet Core Platform architecture, built on the work in [15], [16], [17], [18], and [19]. Figure 1 highlights some key functionalities of the Future Internet Core Platform [20]. Such functionalities can be implemented by means of distributed Agents to be transparently embedded in properly selected network nodes (e.g., Mobile Terminals, Base Stations, Backhaul Network entities, Core Network entities).

The *Sensing and Data Processing functionalities* are in charge of the *monitoring* and the preliminary *filtering* of properly selected possibly heterogeneous information (e.g., parameters describing specific device, network performances, static user profiles, network provider policies, etc.).

The *Cognitive Network Control functionalities* (in the following, also simply referred to as “Network Control functionalities”) consist of a set of cooperative, technology-independent algorithms and procedures which are in charge of the formal description of the Monitored Information in homogeneous metadata, as well as of the proper aggregation of these metadata to form a multi-layer, multi-network *Present Context.* In addition, the Network Control functionalities, on the basis of the Present Context and on the Driving Parameters provided by the *Cognitive Application Interface functionalities* (as explained below), are in charge of taking the control decisions concerning specific Network Control problems.

The *Cognitive Application Interface functionalities* (in the following, also simply referred to as “Cognitive Application Interface”) works directly on the basis of (i) the Present Context, (ii) parameters gathered from the application which can include direct or indirect user feedbacks. The Cognitive Application Interface, among the others, include two key agents, namely the *QoE Evaluator* and the *QoE Controller.*

S. Canale, F. Facchinei, R. Gambuti, L. Palagi and V. Suraci are with the Department of Computer, Control and Management Engineering “Antonio Ruberti” of the University of Rome “La Sapienza”, Rome, Italy (e-mail: {canale,facchinei,gambuti,palagi,suraci}@dis.uniroma1.it).

This work was supported by the Italian Ministry of Education, Research and University project PLATINO, under Grant Agreement n° PON01_01007.

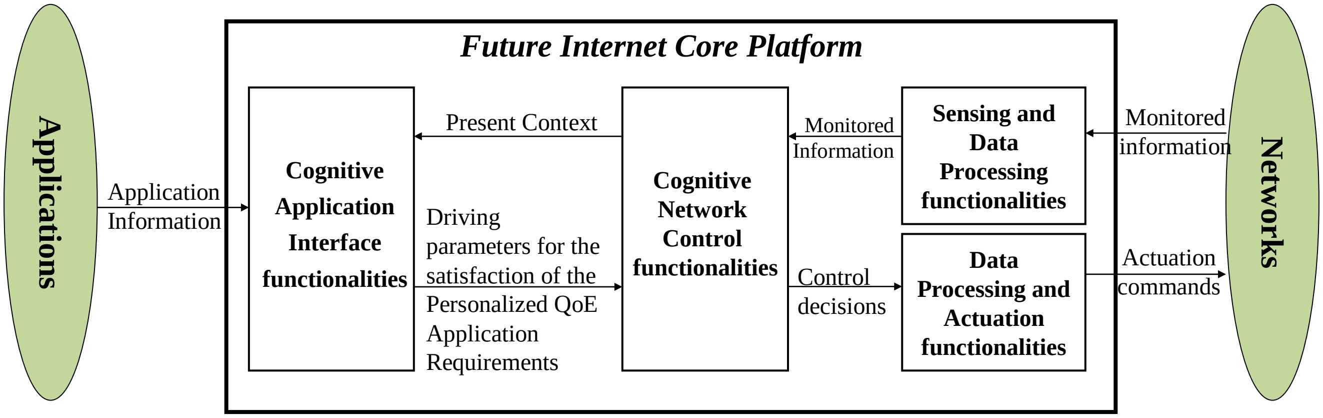

Future Internet Core Platform
Applications
Application Information
Cognitive Application Interface functionalities
Present Context
Driving parameters for the satisfaction of the Personalized QoE Application Requirements
Cognitive Network Control functionalities
Monitored Information
Control decisions
Sensing and Data Processing functionalities
Data Processing and Actuation functionalities
Monitored information
Actuation commands
Networks

Figure 1: Future Internet Core Platform concept

The *QoE Evaluator* is in charge of evaluating, for each in progress application, the personalized QoE expected by the user (hereinafter referred to as *Target QoE*) and the actual, present QoE experienced by the user (hereinafter referred to as *Perceived QoE*).

The *QoE Controller* is in charge of computing, for each in progress application, the Driving Parameters that should drive the Network Control functionalities to take the control decisions aiming at reducing the absolute difference between the Target QoE and the Perceived QoE (namely the so-called *QoE Error*), as well as at optimizing the exploitation of the network resources [27]. The Driving Parameters relevant to a given application can include, among others, QoS aspects (e.g., maximum tolerated delay, minimum throughput to be guaranteed, etc.), security aspects (e.g., desired encryption, allowed network nodes, etc.) or content/service aspects (e.g. the most appropriate service/content mix, etc.). In general, the Driving Parameters are multi-layer and multi-network.

Finally, the *Data Processing and Actuation functionalities* are in charge of "translating" the technology-independent control decisions, taken by the Network Control functionalities, in technology-dependent actuation commands which put into operation on the Networks the above-mentioned decisions.

This paper focuses on the Cognitive Application Interface and, mainly, on the QoE Evaluator. Instead, the Cognitive Network Control functionalities and the QoE Controller are outside the scope of this paper; instances of such functionalities can be found in admission control ([20], [21], and [22]), routing ([23], [24], and [25]), congestion control and scheduling [26], dynamic capacity assignment ([27] and [28]), medium access control [29], and load balancing [30].

III. PRELIMINARY DEFINITIONS

The proposed model for the QoE measuring is based on the dependence of the QoE from the following two types of parameters:

1) the static parameters not varying during the session of use of a particular application; the family of static parameters consists of two main kinds of parameters: the former is related to the service called by the application; the latter to the specific usage of the service in terms of contents, static user profile, invariant parameters of the context relating, for example, to the cloud server that manages the services /contents called by the application;

2) the dynamic parameters that vary in function of time within the session of use of the application; dynamic parameters are classified into the following two sub types:

   a) the Quality of Service (QoS) parameters, typically adopted to objectively measure the QoS; in this paper, the QoS is to be understood in a broad sense, to include not only classical parameters such as delay, packet loss, throughput, but also parameters related to safety and mobility of the users;

   b) the parameters from user feedback used to measure the subjective satisfaction of the users.

Typical example of user feedback is the frequency of the clicks or the selection of satisfaction/dissatisfaction by the user on icons suitably prepared in the application running on the device. Note that such a feedback may be unavailable or be untrustworthy. In this paper, we will refer to the following parameters and notations:

- M is the number of the session typologies that are identified as representative and meaningful of use cases/platform operability;
- for each session typology $i \in \{1,..,M\}$, we denote by h_{ij} the j^{th} session of typology i;
- each session h_{ij} is characterized by a start time $start_{ij}$ and an end time end_{ij};
- for each session h_{ij}, the continuous interval $[start_{ij}, end_{ij}]$ is assumed to be characterized by a finite number of discrete period $\{1,\ldots, k_{ij}\}$;

- we denote by u_{ij} the user requiring the service session h_{ij};
- each user u_{ij} provides a feedback with respect to a specific session j of typology i at time $k \in \{1,\ldots, k_{ij}\}$; we denote by $\phi_{ij}(k) \in H \subseteq R_+$ a suitable function of such a feedback parameter; without loss of generality, we will define $\phi_{ij}(k)$ so to assume values in the interval [0,1]; we will denote by $\phi_{ij}(s,t)$ with $s, t \in\{1,.., k_{ij}\}$ and $s \le t$ the set of feedbacks from user u_{ij} in the subset of discrete times $\{s, s+1,...,t-1, t\} \subseteq \{1,\ldots, k_{ij}\}$, i.e.

$$\phi_{ij}(s,t) = \{ \phi_{ij}(s), \phi_{ij}(s+1), \ldots, \phi_{ij}(t) \}$$

- we denote by U_i the subset of users u_{ij} that has required at least a session of typology i;
- given the function $\phi_{ij}(k)$, we denote by $\rho_i \in H$ the threshold value representing the minimum value of satisfaction of all users in U_i; so that if $\phi_{ij}(k) \ \ge\ \rho_i$ then user u_{ij} is happy in experiencing his session else user u_{ij} is not satisfied;
- $\{p_1,\ldots, p_{q_i}\}$ is the set of q_i clusters in the partition of U_i obtained with respect to the set of user feedbacks $\phi_{ij}(k)$ for k in $\{1,\ldots, k_{ij}\}$; among all possible partitions of U_i, our aim is that of identifying the one such that users in the same group manifested similar feedbacks in similar situations while users in two distinct groups can be considered sufficiently dissimilar from the behavioural point of view;
- given a partition $\{p_1,\ldots, p_{q_i}\}$ of U_i, we will denote by $pr_i(v)$ the most representative element of cluster p_v for $v \in\{1,\ldots, q_i\}$;
- Ψ indicates the finite set of possible function structures, one for each session typology $i \in\{1,..,M\}$, defined starting from both empirical and theoretical results from literature (e.g., [12] in specific case of streaming VOIP);
- f_i indicates the function structure in Ψ associated with the session typology $i \in\{1,..,M\}$; without loss of generality, we will define each f_i such that assumed values range from zero to one ([0,1]);
- $\Omega_{f_{iv}}$ denotes the set of parameters that characterize the function structure f_i for each profile $pr_i(v)$, where $i = 1,\ldots M$ and $v = 1,\ldots q_i$;
- $QoS_{ij}(k)$ indicates the family of measurements of dynamic parameters of QoS during a specific session j of typology i at time $k \in \{1,\ldots, k_{ij}\}$; for sake of readability, even in this case, we will denote by $QoS_{ij}(s,t)$ with $s,t \in\{1,.., k_{ij}\}$ and $s \le t$ the set of measurements of dynamic parameters of QoS during a specific session j of typology i in the subset of discrete times $\{s, s+1,...,t-1, t\} \subseteq \{1,\ldots, k_{ij}\}$, i.e.

$$QoS_{ij}(s,t) = \{ QoS_{ij}(s), QoS_{ij}(s+1), \ldots, QoS_{ij}(t)\})$$

- $QoE_{ij}(k)$ indicates the subjective measure of the level of experience perceived by the user u_{ij} involved at time k in $\{1,\ldots, k_{ij}\}$; in general, the density probability distribution of $QoE_{ij}(k)$ is unknown, nonetheless we can consider a finite set of empirical samples defined by the subjective feedbacks $\phi_{ij}(k)$ coming from the user u_{ij}; from our point of view, each sample $\phi_{ij}(k)$ represents the value assumed by the unknown measure $QoE_{ij}(k)$ for some $k \in \{1,\ldots, k_{ij}\}$

IV. Conceptual framework

The proposed approach for measuring *QoE* as a function of the dynamic *parameters* consists of two main steps. Given:

1) a number M of session typologies corresponding to use cases significant and representative as much as possible; please note that each session concerns one service in the sense that we will assume that the kind of service required by user does not change over the session;

2) for each of the aforementioned session typologies $i \in\{1,..,M\}$, one function structure in a family Ψ (and the related parameters of structure) mostly suitable for *QoE* measurement model, including functional relationship among dynamic parameters of *QoS*;

3) a set U_i of users where each user u_{ij} provides a feedback $\phi_{ij}(k)$ with respect to a specific j^{th} session at time k in $\{1,\ldots, k_{ij}\}$ for some $i \in\{1,..,M\}$; please note that each session h_{ij} concerns one user $u_{ij} = u(h_{ij})$ but one user u_{ij} can of course corresponds to more than one session.

The former step yields a profiling of users in U_i, for each session typology i, that is a partitional clustering $P_i =\{p_1,\ldots, p_{q_i}\}$ of U_i in q_i groups such that users in the same group manifested similar feedbacks in similar situations while users in two distinct groups can be considered sufficiently dissimilar from the behavioural point of view. The aim of the latter step is that of estimating the optimal parameters characterizing the function structure Ψ_i for each session typology $i \in\{1,..,M\}$ and profile $r_{iv} = pr_i(v)$, so to measure the *QoE* of users in p_v on the basis of dynamic parameters of *QoS*. The optimal parameters are identified by analyzing a suitable set of sessions of each typology i in order to automatically recognize interesting correlations between some patterns of dynamic parameters of *QoS* and the feedbacks $\phi_{ij}(k)$ provided by the users u_{ij} in a given group p_v.

V. User profiling

The user profiling module is in charge of collecting and analysing a considerable number of data about each single session of service according to user request and extracting useful information about recurrent patterns naturally arising in the set of analysed sessions. Each session of typology $i \in\{1,..,M\}$ is represented by a *session report* consisting of a complex but structured data record reporting actual values of parameters describing the session, the user requesting the session, the requested service and content, the time series of parameters describing the *QoS*, the actual context and so on. Moreover, the session report records parameters coming directly from user feedback in the form available from the application used by the user to request the specific service.

Typically, user feedback ranges over a finite set of discrete values (e.g. positive integer numbers from 1 to 5). The analysis of the session reports can be conducted in unsupervised way, in the sense that the goal of the analysis is that of identifying groups of similar session reports in order to discriminate two session reports whenever they can be considered sufficiently different. The similarity function adopted to measure the distance between any two session reports plays a crucial role in this kind of applications and depends on the features chosen to describe the time series of parameters reported in each session report. On the basis of this similarity function, each cluster of similar session reports can be defined and identified by the most representative member of the cluster.

Typically, the representative member is the mean or the median point of the cluster and is considered as the *profile* of the users belonging to the cluster. By restricting our attention to time series of parameters describing the $QoS_{ij}(k)$ and the sequence of user feedback $\phi_{ij}(k)$, we can determine q_i groups of users U_i such that the sequences

$$<QoS_{ij}(1,k),\ \phi_{ij}(1,k)> \qquad k=2,..,k_{ij} \qquad (1)$$

are similar for two users in the same group and dissimilar for two users in distinct groups. In this paper, we tackle the sequences $<QoS_{ij}(k),\ \phi_{ij}(k)>$ as multidimensional time series (please see [11]) and, accordingly, clustering similar sequences by means of partitional clustering algorithm (e.g. *k*-means++ [31]).

In order to define a similarity measure to compare two distinct sequences (1) we define the following family of features describing the main characteristics of discrete time series described by the values assumed by the Quality of Service parameters $QoS_{ij}(k)$ and the user feedback $\phi_{ij}(k)$. We consider both features describing the individual characteristics of each single parameter in $QoS_{ij}(k)$ and those describing the combined effects of parameters $QoS_{ij}(k)$ on the user feedback $\phi_{ij}(1,k)$.

In the former family of features, we define the ones measuring the finite derivative ($TQoS_{ij}(k)$) and the discrete integral ($IQoS_{ij}(k)$) functions of each single parameter in $QoS_{ij}(k)$.

$$\mathrm{TQoS_{ij}}(k) = \frac{(\mathrm{QoS_{ij}}(k) - \mathrm{QoS_{ij}}(k-1))}{(k-(k-1))} = \mathrm{QoS_{ij}}(k) - \mathrm{QoS_{ij}}(k-1)$$

$$k \in \{1,\ldots,k\mathrm{ij}\} \qquad (2)$$

$$\mathrm{IQoS_{ij}}(k) = \sum_{s=1}^{k} \mathrm{QoS_{ij}}(s) \qquad k \in \{1,\ldots,k\mathrm{ij}\} \qquad (3)$$

In (2) the first component $TQoS_{ij}(1)$ is evaluated by setting $QoS_{ij}(0) = QoS_0$, where QoS_0 represents the mean value of the parameter $QoS_{ij}(k)$ in $\{1,\ldots,\ k_{ij}\}$. $TQoS_{ij}(k)$ represents the slope of the discrete time series of parameter $QoS_{ij}(k)$ in k. $IQoS_{ij}(k)$ represents the cumulative effect of the discrete time series of parameter $QoS_{ij}(k)$ in k.

In the latter family of features, we consider a more sophisticated feature measuring the *sensitivity* of the user u_{ij} with respect to a given parameter of QoS_{ij}. In order to make the sensitivity as much accurate as possible, we measure the proportional effect of $TQoS_{ij}(k)$ (with respect to a specific parameter $QoS_{ij}(k)$) on the user feedback $\phi_{ij}(k)$ in k and we *weight* this effect by the user reaction time. In fact, we consider that the longer the user reaction time, the less is the sensitivity manifested by the user. In order to evaluate this reaction time, at k we consider the last interval $k'(QoS_{ij})$ when a change in the parameter QoS_{ij} occurred before k:

$$k'(k, QoS_{ij}) := \max\{s \in \{1,..,k\}:\ TQoS_{ij}(s) \neq 0\}$$

Hence, the user reaction time is given by

$$\Delta\, k := k'\big(k, QoS_{ij}\big) - k$$

We define the sensitivity $SQoS_{ij}(k)$ of the user u_{ij} with respect to a given parameter of QoS_{ij} as follows

$$SQoS_{ij}(k) = \frac{\phi_{ij}(k) - \phi_{ij}(k-1)}{TQoSij(k)}\, \Delta k^{-1} \qquad k \in \{1,..,k_{ij}\} \qquad (4)$$

Once the features $TQoS_{ij}(k)$, $IQoS_{ij}(k)$) and $SQoS_{ij}(k)$) describing the finite derivative and the discrete cumulative effect as well as the sensitivity of the user u_{ij} with respect to each single parameter $QoS_{ij}(k)$) have been defined, the clustering algorithm measures the similarity between two distinct sequences (1) by evaluating the distance between the feature vectors defined as

$$[TQoS_{ij}(k), IQoS_{ij}(k), SQoS_{ij(k)}]_{k=1}^{k_{ij}}$$

Given a session of typology $i \in \{1,..,M\}$ and a set U_i of users where each user u_{ij} provides a feedback $\phi_{ij}(k)$ depending on time k with respect to a specific session typology i, the output of the user profiling module is a partitional clustering of users in U_i in q_i groups $\{p_1,\ldots,\ p_{q_i}\}$ such that users in the same group manifested similar feedback in similar situations while users in two distinct groups can be considered sufficiently dissimilar.

VI. QUALITY OF EXPERIENCE MEASUREMENT

Given a session of typology $i \in \{1,..,M\}$ and the partitional clustering of users in U_i where each user u_{ij} in U_i provides a feedback $\phi_{ij}(k)$ in q_i groups $\{p_1,\ldots,\ p_{q_i}\}$, our aim is that of completely identifying a function characterized by a structure f_i in Ψ and a finite set of parameters $\Sigma_{f_{iv}}$ such that

$$QoE_{ijv}(k) = f_i(\Omega_{f_{iv}};\ QoS_{ij}(1,k),\ \phi_{ij}(1,k\text{-}1)) \qquad (5)$$

$$u_{ij} \in p_v,\ v= 1,\ldots, q_i,\ k=2,..,k_{ij}$$

In (5) the unknown measure $QoE_{ij}(k)$ of level of quality experienced by user u_{ij} at time k in session h_{ij} of typology i is estimated by function f_i characterized by parameters $\Omega_{f_{iv}}$ of QoS parameters at time k and previous feedback $\phi_{ij}(k\text{-}1)$ from user u_{ij}. For example, a possible function structure for f is provided by the approximation of the QoE measure by the IQX hypotheses:

$$(\Omega_{f_{iv}};\ QoS_{ij}(1,k),\ \phi_{ij}(1,\ k\text{-}1\ )) = \alpha_{iv}\ e^{\text{-}\beta_{iv}\ QoS_{ij}(k)} + \gamma_{iv} \qquad (6)$$

where $\Omega_{f_{iv}} = \{\alpha_{iv}, \beta_{iv}, \gamma_{iv}\}$ represents the unknown parameters of the structure and the user feedback $\phi_{ij}(1,\ k\text{-}1)$ is not taken into account. In the case of the IQX hypothesis [9], the $QoS_{ij}(1,k)$ function, relatively to a session of typology i, where the user u_{ij} is involved, identifies the trend of a single dynamic parameter of QoS, chosen such that it is as much representative as possible, in respect of the session. Note that one possible way to normalize the function (6), such that it takes values in the interval [0, 1], is to set $\alpha_{iv} = 1$ and $\gamma_{iv} = 0$.

From the algorithmic point of view, given a session of typology i and the partitional clustering of users in a given set U_i of users in q_i groups $\{p_1,\ldots, p_{q_i}\}$ of similar users, we consider a finite set of sequences $<QoS_{ij}(1,k),\ \phi_{ij}(1,k\text{-}1) >$ for u_{ij} in a specific group p_v and estimate the optimal values $\Omega_{f_{iv}}*$ of parameter $\Omega_{f_{iv}}$ of structure f_i in Ψ such that

$$\phi_{ij}(k) = f_i(\Omega_{f_{iv}};\ QoS_{ij}(1,k),\ \phi_{ij}(1,k\text{-}1)) \qquad (7)$$

$$u_{ij} \in p_v,\ v= 1,\ldots, q_i,\ k=2,..,k_{ij}$$

Among all possible feasible solutions to problem (7) we select the one maximizing the generalization capability of f_i. Once the optimal values $\Omega_{f_{iv}}*$ have been estimated by finding a feasible solution to problem (7), $f_i(\Omega_{f_{iv}}*)$ is an approximation of user feedback ϕ_{ij} that we consider as the main indicator of QoE. Therefore, we define the function of measurement of Quality of Experience (QoE^M) as follows

$$QoE^M_{ijv}(k) := f_i(\Omega_{f_{iv}}*;\ QoS_{ij}(1,k),\ \phi_{ij}(1,k\text{-}1)) \qquad (8)$$

$$u_{ij} \in p_v,\ v= 1,\ldots, q_i,\ k=2,..,k_{ij}$$

The personalized QoE expected by the user (the so called *Target QoE*) is directly provided by the QoE^M_{ijv} so to have a reference value for the QoE expected by the user u_{ij} by means of his user profile p_v. Once a user u_{ij} is assigned the Target QoE QoE^M_{ijv}, for every new session the actual, present QoE experienced by the user (the so called *Perceived QoE*) will be captured on line in the form of user feedback ϕ_{ij}. The *QoE Controller* is now in position of computing the Driving Parameters that should drive the Network Control functionalities to take the control decisions aiming at reducing the absolute difference between the Target QoE and the Perceived QoE (namely the so-called *QoE Error*), as well as at optimizing the exploitation of the network resources (cf. Figure 1).

VII. Conclusions

In this paper we defined the problem of QoE evaluation as a user profile based procedure. We proposed a QoE evaluation framework that allows to automatically identify distinct user profiles on the basis of subjective user feedbacks with respect to a generic service by extracting suitable features and by adopting off the shelf clustering algorithms. The framework is independent on the specific service offered by the user, as well as on the specific QoS parameters. It represents a fundamental step for providing the fundamental reference for the QoE Controller module by realizing a complete personalization of the QoS/QoE correlation.

Acknowledgments

The authors wish to thank all the members of the FI-WARE and PLATINO consortia, and, in particular, F. Delli Priscoli and A. Pietrabissa for their valuable contributions; nevertheless, the concepts and the results contained in this paper are not necessarily the ones of the consortia.

References

[1] M.S. Musthaq, B Augustin, and A.Mellouk, "Empirical Study based on Machine Learning Approach to Asses the Qos/qoe Correlation*", Proc. 17th European Conf. Networks and Optimal Comm.,pp. 1-7*, 2012.

[2] T. Zinner, O. Hohlfeld, O. Abboud, and T. Hoßfeld "Impact of Frame Rate and Resolution on Objective QoE Metrics", *International Workshop on Quality of Multimedia Experience*, 2010, Trondheim, June 2010.

[3] Herman, H., Rahman, A. A., Syahbana, Y. A., & Bakar, K. A. "Nonlinearity modelling of QoE for video streaming over wireless and mobile network", *Second international conference on intelligent systems, modelling and simulation (ISMS)*, 2011, pp. 313–317.

[4] Oliver Hohlfeld, Rüdiger Geib, Gerhard Haßlinger, "Packet loss in real-time services: Markovian models generating QoE impairments", *Proc. of the 16th International Workshop on Quality of Service, IWQoS*, June 2008, pp. 239–248.

[5] E. Dillon, G. Power, M. O. Ramos, M. A. Callejo Rodríguez, J. R. Argente, M. Fiedler, D. S. Tonesi, "PERIMETER: A Quality of Experience Framework", *FIS 2009: Future Internet Symposium 2009*, Berlin (Germany), Sep. 1-3 2009.

[6] H. A. Tran and A. Mellouk, "Qoe model driven for network services", *Wired/Wireless Internet Communications*, pp. 264–277, 2010.

[7] F. Neves, S. Cardeal, S. Soares, P. Assuncao, and F. Travares, "Quality model for monitoring QoE in VOIP services", *in EUROCON – International Conference on Computer as a Tool (EUROCON)*, 2011 IEEE, April 2011, pp. 1-4.

[8] Baraković, S. and Skorin-Kapov, L., "Survey and Challenges of QoE Management Issues in Wireless Networks", *Journal of Computer Networks and Communications 2013*, 28.

[9] T. Hoßfeld, D. Hock, P. Tran-Gia, K. Tutschku, M. Fiedler, "Testing the IQX Hypothesis for Exponential Interdependency between QoS and QoE of Voice Codecs iLBC and G.711", *18th ITC Specialist Seminar on Quality of Experience*, Karlskrona, Sweden, May 2008.

[10] Laghari, K., Crespi, N., Connelly, K., “Toward total quality of experience: A QoE model in a communication ecosystem”, *IEEE Communications Magazine 50(4)*,2012, pp.58–65.

[11] D. Kotsakos, G. Trajcevski, D. Gunopulos, C. Aggarwal, “Time-series Data Clustering”, *in Data Clustering: Algorithms and Application*, CRC Press, 2013.

[12] M. Fiedler, T. Hossfeld, P. Tran-Gia. “A generic quantitative relationship between Quality of Experience and Quality of Service”, *IEEE Network*, vol. 24, no. 2, Mar./Apr. 2010.

[13] FI-WARE (Future Internet Core Platform) EU FP7 project, contract n. 285248, www.fi-ware.eu;

[14] PLATINO (Grant Agreement n° PON01_01007).

[15] Delli Priscoli, F., "A Fully Cognitive Approach for Future Internet", Special Issue on "Future Network Architectures" of "Future Internet", Molecular Diversity Preservation International (MDPI), Vol. 2, January 2010, pp. 16-29.

[16] M. Castrucci, F. Delli Priscoli, A. Pietrabissa, V. Suraci, “A Cognitive Future Internet Architecture”, The Future Internet Future Internet Assembly 2011, Springer Berlin/Heidelberg (DE), Lecture Notes in Computer Science, Vol. 6656, May 2011, doi: 10.1007/978-3-642-20898-0_7.

[17] M. Castrucci, M. Cecchi, F. Delli Priscoli, L. Fogliati, P. Garino, V. Suraci, "Key Concepts for the Future Internet Architecture", Future Network & Mobile Summit 2011, Warsaw, 15-17 June 2011

[18] Delli Priscoli, F., Castrucci, M., “A proposal for future internet architecture”, 2010 Future Network and Mobile Summit, 16-18 June 2010, Florence,

[19] F. Delli Priscoli, L. Fogliati, A. Palo, A. Pietrabissa, "Dynamic Class of Service Mapping for Quality of Experience Control in Future Networks", World Telecommunication Congress (WTC), Berlin, June 2014.

[20] F. Delli Priscoli, V. Suraci, M. Castrucci, “Cognitive Architecture for the Internet of the Future”. 6th international workshop on next generation networking middleware. October 26-30, 2009. ISBN 978-3-930736-14-0. Pag. 105-112.

[21] A. Pietrabissa, "Admission Control in UMTS Networks based on Approximate Dynamic Programming", European Journal of Control (The European Union Control Association, Lavoisier, France), Vol. 14, N. 1, pp. 62-75 , January 2008, DOI:10.3166/ejc.14.62-75.

[22] A. Pietrabissa, “An Alternative LP Formulation of the Admission Control Problem in Multi-Class Networks”, IEEE Transaction on Automatic Control, (IEEE Control System Society, USA), Vol. 53, N. 3, pp. 839-845, April 2008, DOI: 10.1109/TAC.2008.919516.

[23] C. Bruni, F. Delli Priscoli, G. Koch, A. Pietrabissa, L. Pimpinella, “Network decomposition and multi-path routing optimal control” , Transactions on Emerging Telecommunications Technologies (John Wiley & Sons, Inc., USA), Vol. 24, pp. 154-165, 2013, published on line 2 May 2012, doi: 10.1002/ett.2536.

[24] Oddi, G., Pietrabissa, A. "A distributed multi-path algorithm for wireless ad-hoc networks based on Wardrop routing," *Proc. 21st Mediterranean Conference on Control and Automation (MED 2013)*, June 25-28, 2013, Platanias-Chania, Crete, Greece, pp. 930-935, doi: 10.1109/MED.2013.6608833.

[25] Manfredi, S. “A Reliable and Energy Efficient Cooperative Routing Algorithm for Wireless Monitoring Systems”, IET Wireless Sensor Systems 2 (2012), 128-135.

[26] Delli Priscoli, F., Isidori, A., “A control-engineering approach to integrated congestion control and scheduling in wireless local area networks”, (2005) Control Engineering Practice, 13 (5), pp. 541-558, doi: 10.1016/j.conengprac.2004.04.016.

[27] F. Delli Priscoli, A. Pietrabissa, “Design of a bandwidth-on-demand (BoD) protocol for satellite networks modeled as time-delay systems”, Automatica, International Federation of Automatic Control (Elsevier, Great Britain), Vol. 40, Issue 5, pp. 729-741, May, 2004, DOI:10.1016/j.automatica.2003.12.013.

[28] R. Cusani, F. Delli Priscoli, G. Ferrari, M. Torregiani, "A Novel MAC and Scheduling Strategy to Guarantee QoS for the New-Generation Wireless LAN", Special Issue on "Mobile and Wireless Internet: Architecture and Protocols" of IEEE Wireless Communications (IEEE Personal Communications), IEEE's Computer and Vehicular Technology Societies (U.S.A.), Issue 3, June 2002, pp. 46-56.

[29] Francesco Delli Priscoli, "A Control Based Solution for Integrated Dynamic Capacity Assignment, Congestion Control and Scheduling in Wireless Networks", European Journal of Control, European Union Control Association (EUCA), Issue n.2/2010, pp. 169-184.

[30] Macone D., Oddi G., Palo A., Suraci V., "A Dynamic Load Balancing Algorithm for Quality of Service and Mobility Management in Next Generation Home Networks", Telecommun Syst (Springer), 2013, DOI 10.1007/s11235-013-9697-y.

[31] Arthur D., Vassilvitskii S., k-means++: the advantages of carefull seeding, in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 1027-1035, 2007.

Social Relevance Feedback: an Innovative Scheme Based on Multimedia Content Power

Klimis S. Ntalianis, and Anastasios D. Doulamis

***Abstract*—**In this paper, a novel relevance feedback algorithm based on multimedia content power is proposed. The iterative query-by-example algorithm recursively estimates the similarity measure, which is used for data ranking. To do so, it uses a set of relevant/irrelevant samples feedbacked by the user to the system so that the adjusted response is a better approximation of the current user's information needs and preferences. In particular, using concepts of functional analysis, the similarity measure is expressed as a parametric form of known monotone increasing functional components. Another innovation of the proposed scheme is the definition of multimedia content power, a concept that tries to quantify the degree of influence of a multimedia file on users. When a user performs an action to a particular file (such as writing a comment, sharing etc), it indicates the file has some influence on her. The iterative relevance feedback algorithm takes into consideration both the visual characteristics of a multimedia file and its power (influence). Experimental results exhibit the advantages of the proposed scheme and interesting directions are discussed for future work.

***Keywords*—**Multimedia Retrieval, Social Media, Relevance Feedback (RF), Multimedia Content Power, Social Computing

I. Introduction

RESEARCH on social media has greatly intensified in the past few years given the significant interest from the application's perspective and the associated unique technical and social science challenges and opportunities. This research agenda is multidisciplinary in nature and has drawn attention from research communities in all major disciplines. From an information technology standpoint, social media research has primarily focused on social media analytics and, more recently, social media intelligence. Social media analytics is concerned with developing and evaluating informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data, usually driven by specific requirements from a target application. From a technical perspective, social media analytics research faces several unique challenges. First, social media contains an enriched set of data or metadata, which have not been treated systematically in data- and text-mining literature. Examples include tags (annotations or labels using free-form keywords); user-expressed subjective opinions, insights, evaluation, and perspectives; ratings; user profiles; and both explicit and implicit social networks. Second, social media applications are a prominent example of human-centered computing with their own unique emphasis on social interactions among users. Hence, issues such as context-dependent user profiling and needs elicitation as well as various kinds of human-computer interaction considerations must be reexamined. Third, social media data are dynamic streams, with their volume rapidly increasing. The dynamic nature of such data and their sheer size pose significant challenges to computing in general and to semantic computing in particular [1]. Thus each file posted on social media is not characterized only by its content but also by its context.

On the other hand social media content has already been used in a variety of applications such as for the ranking of news stories [2], for the profiling of user preferences [3] and even for products' recommendations [4]. However this type of conversational, user-generated content found in social media, might be used to add value to more traditional media, such as images and video. Towards this direction and in order to facilitate information search in social media, this paper proposes an image retrieval mechanism, based on a social interactions. In particular we focus on the problem of relevant image retrieval, posted by our social media friends. This can be accomplished by relevance feedback and several schemes have been proposed in the literature.

In Navigation-Pattern-based Relevance Feedback (NPRF) [5] the feedback iterations are reduced substantially by using the navigation patterns discovered from the user query log. In terms of effectiveness, NPRF makes use of the discovered navigation patterns and three kinds of query refinement strategies: Query Point Movement (QPM), Query Reweighting (QR), and Query Expansion (QEX), to converge to the desired content. In [6] graphs have been used to represent images, transforming the region correspondence estimation problem into an inexact graph matching problem. An optimization technique has also been proposed to find the solution. In [7] images were represented by low-level visual features. A mapping has been designed so that the effective subspace is selected for separating positive samples from negative samples based on a number of observations. Furthermore the Biased Discriminative Euclidean Embedding (BDEE) has been proposed which parameterizes samples in the original high-

K. S. Ntalianis is with the Technological Educational Institute of Athens, Department of Marketing – Online Computing Group, 12210, Egaleo, Athens, GREECE, e-mail: kdal75@gmail.com

A. D. Doulamis is with the National Technical University of Athens, Department of Rural and Surveying Engineering, Athens, GREECE, e-mail: adoulam@cs.ntua.gr

dimensional ambient space to discover the intrinsic coordinate of image low-level visual features. Pinview is proposed in [8]. Its main difference to the literature is that it is based on an implicit relevance feedback mechanism that is activated during each search session. Another interesting scheme is presented in [9]. In this case multinomial relevance feedback is introduced and system knowledge is modeled by using a Dirichlet process. Finally in [10] a multilayer neural network is used in connection with a Fuzzy Radial Basis Function Network (FRBFN), to enhance relevance feedback during the retrieval cycle. The FRBFN tries to introduce human decision fuzziness into the system, while a fast learning algorithm (without back-propagation) is applied to reduce the convergence time.

Even though interesting, the aforementioned approaches consider a constant type of similarity measure, e.g., the correlation criterion, regulating only the importance of extracted descriptors to the similarity metric, instead of the similarity type itself. Additionally they do not take into consideration content's context especially in case of social networks were much more information is provided (likes, comments, sharing etc.). On the contrary this paper introduces two significant innovations: (a) the problem of relevance feedback is addressed in the most generic form by optimally updating the similarity measure type to the current users' information needs and preferences. In this case, instead of adjusting the degree of importance of each descriptor, the similarity measure itself is estimated through an on-line efficient and recursive learning strategy. Therefore, no restrictions on the type of the similarity are imposed (b) a new metric is proposed, namely Multimedia Content Power (MCP), aiming at estimating the importance of each posted multimedia file based on its influence to the audience. The more attention a file receives the more influential it can be considered. Thus for each file an MCP value is calculated. These novelties are combined to produce a social relevance feedback mechanism. In particular, the algorithm is based on a similarity measure modeling and estimation strategy. Initially the similarity measure is assumed to be of any non-linear function type. In the following, the similarity is modeled as a non-linear parametric relation of known functional components [24], [25], using concepts derived from functional analysis. The contribution of each component to the similarity measure is estimated based on a set of selected relevant / irrelevant samples interactively provided by the user to express the current information needs. As a result, at each feedback iteration, the type of similarity measure is estimated resulting in a generalized non-linear relevance feedback scheme. The contribution of each functional component to the similarity measure is recursively estimated through an efficient on-line learning strategy. Furthermore the relevance function contains the MCP value of each file. Experimental results on Facebook's content illustrate the promising performance of the proposed scheme.

The rest of the paper is organized as follows: In Section 2 the visual content representation methodology is described together with visual content relevance metrics. Section 3 focuses on similarity measure modeling and estimation, while in Section 4 the relevance feedback algorithm based on MCP is analyzed. Results on real-world content are provided in Section 5. Finally Section 6 concludes this paper.

II. Visual Content Representation & Relevance

A. *Visual Content Representation*

Two of the most important aspects of each relevance feedback algorithm are: (a) visual representation of each file (i.e. image, video) under consideration and (b) relevance estimation between two files.

In order to provide efficient visual content representation, a semantic high-level description of files is required. Several attempts have been carried out in the literature, including MPEG-7 [11], the European COST211ter group [12], SIFT [13] and SURF [14]. In our case global and object-based descriptors are used [15]. The first refer to global visual characteristics, while the second exploit region-based properties, obtained by applying a segmentation algorithm. As global descriptors of still images, the global color and texture are used. For the estimation of the object-based descriptors, initially a segmentation algorithm is applied, which exploits color information. In our case, a multiresolution implementation of the Recursive Shortest Spanning Tree algorithm (RSST) is adopted to perform the segmentation task due to its efficiency and low computational complexity [16]. For each color segment, the average color, size and segment location are extracted as appropriate descriptors. Afterwards, both global and object-based descriptors are organized in a fuzzy classification scheme described in [17], where each file is represented by a vector. Fuzzy organization, apart from reducing possible noise in the extracted descriptors and uncertainty in their values due to quantization errors, interference problems and inaccuracies/instabilities, it also provides a physical interpretation of the visual content, which is closer to human perception.

B. *Visual Content Relevance*

After representing each file by a vector, next, a way for estimating relevance between files is needed. Towards this direction let us assume a query by example operation for content-based retrieval. This means that the user provides queries in the form of images, for which similar analysis is performed producing a feature vector. Now in order to find similarity between two files (represented by vectors), different metrics can be used. The most common is the Euclidean distance, where in its generalized form it is defined as [18]:

$$d(\mathbf{f}_q, \mathbf{f}_i) = (\mathbf{f}_q, \mathbf{f}_i)^T \cdot \mathbf{W} \cdot (\mathbf{f}_q, \mathbf{f}_i) \tag{1}$$

where $\mathbf{f}_q$ is the feature vector of the query and $\mathbf{f}_i$ the feature vector of a sample. **W** is a real symmetric matrix which contains the weights that regulate the degree of importance of the feature elements to the similarity measure. In case that no interconnection among different feature elements is permitted, matrix **W** becomes diagonal and the resulted similarity measure is called Weighted Euclidean Distance.

Another interesting similarity measure is the cross-correlation criterion, which indicates how similar two feature vectors are and thus provides a measure for their content similarity:

$$\rho_{\mathbf{w}}(\mathbf{f}_q,\mathbf{f}_i) = \frac{\sum_{k=1}^{P} w_k \cdot f_{q,k} \cdot f_{i,k}}{\sqrt{\sum_{k=1}^{P} w_k^2 \cdot f_{q,k}^2} \cdot \sqrt{\sum_{k=1}^{P} f_{i,k}^2}} \qquad (2)$$

where $f_{q,k}$ and $f_{i,k}$ are the k$^{\text{th}}$ element of vectors $\mathbf{f}_q$ and $\mathbf{f}_i$ respectively. The variable P in indicates the size of the feature vector, while parameters w_k the relevance of the k$^{\text{th}}$ element of the query feature vector to the selected ones. In order to accomplish a generic non-linear relevance feedback scheme, in this paper the type of similarity measure is dynamically estimated. In particular, the similarity measure, say $d(\cdot)$, is modeled as a continuous function $g(\cdot)$ of the difference between the query feature vector $\mathbf{f}_q$ and the feature vector $\mathbf{f}_i$ of the i$^{\text{th}}$ sample in the database:

$$d(\mathbf{f}_q, \mathbf{f}_i) = g(\mathbf{f}_q - \mathbf{f}_i) \qquad (3)$$

Equation (3) models any non-linear similarity measure of the query feature vector $\mathbf{f}_q$ and the i$^{\text{th}}$ sample in the database $\mathbf{f}_i$ so that the current user's information needs and preferences are satisfied as much as possible. However in equation (3), the feature vectors $\mathbf{f}_q$ and $\mathbf{f}_i$ are involved, which affect the retrieval results. For this reason in this paper, we concentrate on the estimation of the most appropriate similarity measure $g(\cdot)$, which satisfies the user information needs and preferences as much as possible for a *given feature vector representation*.

III. Similarity Measure Modeling & Estimation

A. Similarity Measure Type Modeling

Since $g(\cdot)$ is unknown, initially modeling of $g(\cdot)$ is performed in a parametric form. Towards this direction, it can be proved that any continuous non-linear function can be expressed as a parametric relation of known functional components $\mathbf{\Phi}_l\ (\cdot)$ within any degree of accuracy [19]. In our case, we have that:

$$d(\mathbf{f}_q,\mathbf{f}_i) = g(\mathbf{f}_q - \mathbf{f}_i) \approx \sum_{l=1}^{L} v_l \cdot \Phi_l\left(\sum_{k=1}^{P} w_{k,l} \cdot (f_{q,k} - f_{i,k})\right) \qquad (4)$$

where L expresses the approximation order of function $g(\cdot)$. In equation (4), v_l and $w_{k,l}$ correspond to model parameters, while the $\mathbf{\Phi}_l\ (\cdot)$ to functional components. It is clear that the approximation precision increases, as the order L increases.

The most familiar class of functional components $\mathbf{\Phi}_l\ (\cdot)$ is the sigmoid functions:

$$\mathbf{\Phi}_l\ (x) = 1/(1\text{-}\exp(\ \text{-}a\cdot x)) \qquad (5)$$

where a is a constant which regulates the curve steepness. It should be mentioned that the parameters v_l, $w_{k,l}$ are not related to the weighted factors (degree of importance) of the descriptors, which are used other relevance feedback approaches. On the contrary, they express the coefficients (model parameters) on which function $g(\cdot)$ is expanded to the respective functional components.

From equation (4), it seems that $P \times L$ parameters are required to approximate any continuous similarity measure of order L. The number of parameters can be reduced by imposing constraints on v_l and $w_{k,l}$, which, however, restrict the type of similarity measure of (4). Particularly, let us assume that the same parameters are assigned to all functional components $\mathbf{\Phi}_l\ (\cdot)$. This means that $w_{k,l} = w_{k,q}\ \forall l,\ q$ and therefore $w_{k,l} = w_k$ since they depend only on the feature vector index and not on the index of the functional components. If we further assume that the functional components are linear, we conclude that:

$$d(\mathbf{f}_q,\mathbf{f}_i) = \left(\sum_{l=1}^{L} v_l\right) \cdot \sum_{k=1}^{P} w_k \cdot (f_{q,k} - f_{i,k}) \qquad (6)$$

which simulates the weighted Euclidean distance with free parameters the P variables w_k. In this case, the parameters v_l do not affect the performance of the similarity measure since they just multiply the overall similarity.

B. Similarity Measure Type Estimation

Using, equation (4), it is clear that the estimation of the similarity measure is equivalent to the estimation of coefficients v_l, $w_{k,l}$. In particular, let us denote as $S^{(r)}$ a set, which contains selected relevant/ irrelevant samples at the r feedback iteration of the algorithm. Set $S^{(r)}$ is of the form:

$$S^{(r)} = \{\ldots,(\mathbf{f}_q - \mathbf{f}_i, R_i,\ldots\} = \{\ldots,(\mathbf{e}_i, R_i),\ldots\} \qquad (8)$$

where R_i expresses the respective degree of relevance. Negative/positive values of R_i express irrelevant/relevant. Let us denote as $v_l(r+1)$, $w_{k,l}(r+1)$ the model parameters at the $(r{+}1)$ feedback iteration. These coefficients are estimated so that, the similarity measure, after the r$^{\text{th}}$ feedback iteration, equals to the degree of relevance assigned by the user over all selected data:

$$d^{(r+1)}(\mathbf{f}_q,\mathbf{f}_i) = g^{(r+1)}(\mathbf{f}_q - \mathbf{f}_i) =$$
$$= \sum_{l=1}^{L} v_l(r+1) \cdot \Phi_l\left(\sum_{k=1}^{P} w_{k,l}(r+1) \cdot (f_{q,k} - f_{i,k})\right) \approx \qquad (9)$$
$$\approx R_i,\ \ with\ \ i \in S^{(r)}$$

where $d^{(r+1)}\ (\cdot)$ expresses the non-linear similarity measure at the $(r{+}1)^{\text{th}}$ feedback iteration of the algorithm.

Usually, the number of samples of set $S^{(r)}$ at the r$^{\text{th}}$ feedback iteration is much smaller than the number of coefficients v_l, $w_{k,l}$ that should be estimated. Therefore, equation (9) is not sufficient to uniquely identify parameters v_l, $w_{k,l}$. Uniqueness is achieved by an additional requirement, which takes into consideration the variation of the similarity measure. In particular, among all possible solutions, the one that satisfies (9) and simultaneously causes a minimal modification of the already estimated similarity measure is selected as the most appropriate.

Let us denote as S a set, which contains relevant/irrelevant samples with respect to several queries. Set S is used for the initial estimation of the similarity measure based on a least squared minimization algorithm [20]. At each feedback iteration, the set S is augmented by adding new selected relevant/irrelevant samples. In order to retain a constant size of S, the older samples are removed from S as new samples are added. Then, the requirement of minimal modification of the already estimated similarity measure is expressed as:

$$\min imize\, B(r) = \left\| E^{(r+1)} - E^{(r)} \right\|_2 \tag{10}$$

where $E^{(r)} = \frac{1}{2}\cdot\sum_{i\in S}\{g^{(r)}(\mathbf{f}_q - \mathbf{f}_i) - R_i\}^2$ corresponds to the error of the similarity measure over all data of S at the r$^{\text{th}}$ feedback iteration. As a result the model parameters of the similarity measure are estimated by the following constraint minimization problem:

$$\min imize\, B(r) = \left\| E^{(r+1)} - E^{(r)} \right\|_2 \tag{11a}$$

subject to

$$\begin{aligned} & d^{(r+1)}(\mathbf{f}_q,\mathbf{f}_i) = g^{(r+1)}(\mathbf{f}_q - \mathbf{f}_i) = \\ & = \sum_{l=1}^{L} v_l(r+1)\cdot\Phi_l\left(\sum_{k=1}^{P} w_{k,l}(r+1)\cdot(f_{q,k} - f_{i,k})\right) \approx \\ & \approx R_i,\ \ with\ (f_{q,k} - f_{i,k}) \in S^{(r)} \end{aligned} \tag{11b}$$

The constrained minimization of (11) is performed using the recursive algorithm proposed in [20].

IV. Relevance Feedback Based on Multimedia Content Power

As stated in the Introduction section, each multimedia file is not characterized only by its content but also by its context [21]. More specifically in this paper it is assumed that a file which has received more attention may be more important than a visually similar file that has received less attention. In social media attention is explicitly expressed by liking, sharing or commenting on a post.

Towards this direction let us define the Multimedia Content Power (MCP) as a degree of influence of a multimedia file on other users in social media. When a user performs an action to a particular file (such as writing a comment), it indicates the file has some influence on the user. This observation leads to a new method of quantifying the *MCP*. The main idea is that the *MCP* can be computed by adding up the weighted frequencies of other users' activities induced by that particular file. A multimedia file can have either direct influence on other users who access the file directly (e.g. find it outside facebook and post it on their walls), or indirect influence on others who access the file in someone else's wall. In this paper, we call the former as DirectMultimediaContentPower (*DMCP*) and the latter IndirectMultimediaContentPower (*IMCP*). The *MCP* of a multimedia file can be the sum of the weighted values of *DMCP* and *IMCP*:

$$MCP(MF_{i,j}) = w_{dir} \times DMCP(MF_{i,j}) + w_{ind} \times IMCP(MF_{i,j}) \tag{12}$$

where $MF_{i,j}$ is the jth multimedia file of the ith user of a social network (U_i), while w_{dir} and w_{ind} change the degree of relative importance of *DMCP* and *IMCP*. In this paper, in order to compute the content power of a multimedia file efficiently, we use a diffusion history structure which stores the activities done on the file and the history of diffusion of each original file.

In particular for each posted multimedia file $MF_{i,j}$, three vectors are defined, $\mathbf{l}_{MF_{i,j}}$, $\mathbf{p}_{MF_{i,j}}$ and $\mathbf{c}_{MF_{i,j}}$, corresponding to likes, shares and comments the multimedia item has received respectively:

$$\mathbf{l}_{MF_{i,j}} = [l^i_{F^1}, l^i_{F^2}, \dots, l^i_{F^M}, l^i_{F^{M+1}}] \tag{13a}$$

$$\mathbf{p}_{MF_{i,j}} = [p^i_{F^1}, p^i_{F^2}, \dots, p^i_{F^M}, p^i_{F^{M+1}}] \tag{13b}$$

$$\mathbf{c}_{MF_{i,j}} = [c^i_{F^1}, c^i_{F^2}, \dots, c^i_{F^M}, c^i_{F^{M+1}}] \tag{13c}$$

where $l^i_{F^1}$ equals to 1 if the first friend of U_i (denoted by $F^1_{U_i}$) has liked the respective multimedia item and equals to 0 if friend $F^1_{U_i}$ has not liked the respective multimedia item. Similarly, $p^i_{F^1}$ equals to 1 if friend $F^1_{U_i}$ has shared the respective multimedia item and equals to 0 if friend $F^1_{U_i}$ has not shared the respective multimedia item. Variable $c^i_{F^1}$ equals to the number of comments friend $F^1_{U_i}$ has made to the respective multimedia item, while $l^i_{F^{M+1}}$, $p^i_{F^{M+1}}$ and $c^i_{F^{M+1}}$ are used to count the likes, shares and comments the multimedia item has received from users that are not friends with U_i. Vectors of Eqs. 13(a) – 13(c) take into consideration also the respective friend that has interacted with a file. This is due to the fact that maybe different weights can be used to evaluate the importance of each file according to the importance of each friend. Finally we denote as $L_{MF_{i,j}}$, $P_{MF_{i,j}}$ and $C_{MF_{i,j}}$ three variables that count the total number of likes, shares and comments a multimedia file has received respectively:

$$L_{MF_{i,j}} = \sum_{r=1}^{M+1} l^i_{F^r} \tag{14a}$$

$$P_{MF_{i,j}} = \sum_{r=1}^{M+1} p^i_{F^r} \tag{14b}$$

$$C_{MF_{i,j}} = \sum_{r=1}^{M+1} c^i_{F^r} \tag{14c}$$

Then *IMCP* can be calculated as:

$$IMCP(MF_{i,j}) = L_{MF_{i,j}} + P_{MF_{i,j}} + C_{MF_{i,j}} \tag{15}$$

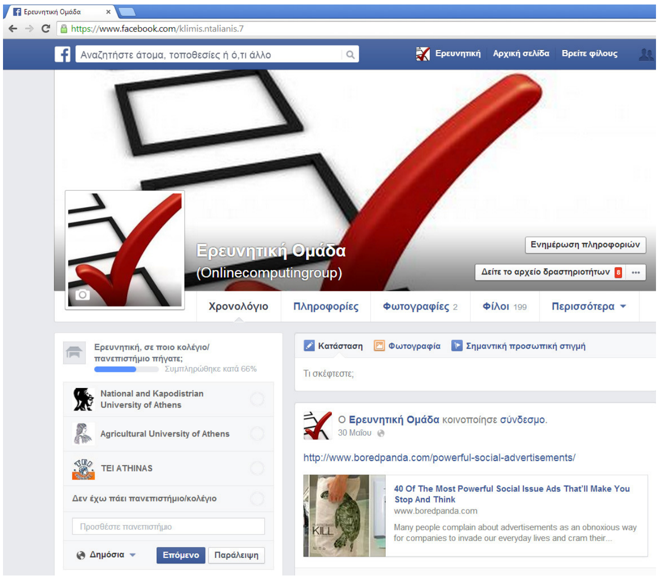
Additionally $DMCP(MF_{i,j})$ equals to the number of times a multimedia file has been posted to walls from its original source (not shared from another wall).

Now in order to consider *MCP* values during feedback, let us recall that $\mathbf{f}_q$ is the feature vector of the query and $\mathbf{f}_i$ the feature vector of a sample. Then, for simplicity reasons, we can denote by $MCP_{\mathbf{f}_i}$ the *MCP* value of the file corresponding to vector $\mathbf{f}_i$. Then, by modifying Eq.(6), the RF algorithm that considers multimedia content power can be based on the following equation:

$$d'(\mathbf{f}_q,\mathbf{f}_i) = w_{MCP}\cdot\frac{1}{MCP_{\mathbf{f}_i}}\cdot\left(\left(\sum_{l=1}^{L} v_l\right)\cdot\sum_{k=1}^{P} w_k\cdot(f_{q,k}-f_{i,k})\right)+ \qquad (16)$$

$$+\,w_{VR}\cdot\left(\left(\sum_{l=1}^{L} v_l\right)\cdot\sum_{k=1}^{P} w_k\cdot(f_{q,k}-f_{i,k})\right)$$

The above Eq. (16) takes advantage of the importance of each multimedia file to calculate its distance from other files. In case that $MCP_{\mathbf{f}_i}$ is large, the distance between the two files decreases. On the other hand if $MCP_{\mathbf{f}_i}$ is small, the distance between the two files increases. In order to avoid neutralizing the real visual distance between two files, weights w_{MCP} and w_{VR} are used with $w_{MCP} + w_{VR} = 1$. The largest the w_{MCP} the less the importance to the real content that is retrieved and the more the importance to the context.



Ερευνητική Ομάδα
(Onlinecomputingroup)

Ερευνητική, σε ποιο κολέγιο/πανεπιστήμιο πήγατε;
Συμπληρώθηκε κατά 66%

National and Kapodistrian University of Athens

Agricultural University of Athens

TEI ATHINAS

Δεν έχω πάει πανεπιστήμιο/κολέγιο

Τι σκέφτεστε;

Ο Ερευνητική Ομάδα κοινοποίησε σύνδεσμο.
30 Μαΐου

http://www.boredpanda.com/powerful-social-advertisements/

40 Of The Most Powerful Social Issue Ads That'll Make You Stop And Think
www.boredpanda.com
Many people complain about advertisements as an obnoxious way for companies to invade our everyday lives and cram their...

Figure 1: The first page of the Online Computing Group

V. Experimental Results

In order to evaluate the performance of the proposed social relevance feedback scheme, we have setup an experimentation phase on real world images, posted on Facebook. In particular we have recorded the Facebook walls' information of 100 members of the Online Computing Group (Onlog) (www.facebook.com/klimis.ntalianis.7) (Figure 1) for a period of three months, discarding all other posts except of images. In total 372,567 images have been gathered in the database. However, some of them appeared two (34,887), three (17,912), four (8,517), five (2,981) or more (872) times. After discarding multiple copies, 259,861 unique images have remained, meaning that only about 69.7 of posted images were unique. After discarding duplicates, for each unique image a feature vector $\mathbf{f}_i$ was formulated using the method proposed in [17]. Furthermore *DMCP* was calculated by counting the times one image was posted on a user's wall by its original source. Additionally *IMCP* was calculated by Eq. (15) (aggregation of likes, shares, comments). Then an *MCP* value for each image was estimated by Eq. (12). In the current results w_{dir} was set equal to 1.5 while w_{ind} was set equal to 0.9, meaning that direct posts from the original source were strengthen, while indirect actions were weaken. This is due to the fact that when someone posts an image from its original source it probably means that she was influenced. On the other hand likes and comments several times have to do with the user and not with the content itself (e.g. sometimes when user A likes user B, user A may like/comment on the content that user B posts).

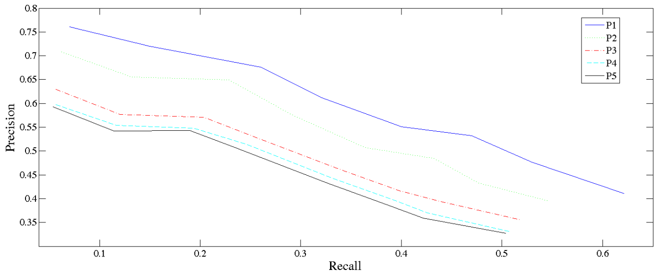


Figure 2: Average precision-recall curve for 1,000 randomly submitted queries and for all five pairs (P_1 – P_5).

Next we examined the quality of the proposed social relevance feedback algorithm. Initially we chose five pairs of w_{MCP} and w_{VR} P_1: (0.1, 0.9), P_2: (0.2, 0.8), P_3: (0.3, 0.7), P_4: (0.4, 0.6) and P_5: (0.5, 0.5). For each pair we submitted 1,000 randomly selected query images to the database and obtained the average precision-recall curve (Figure 2).

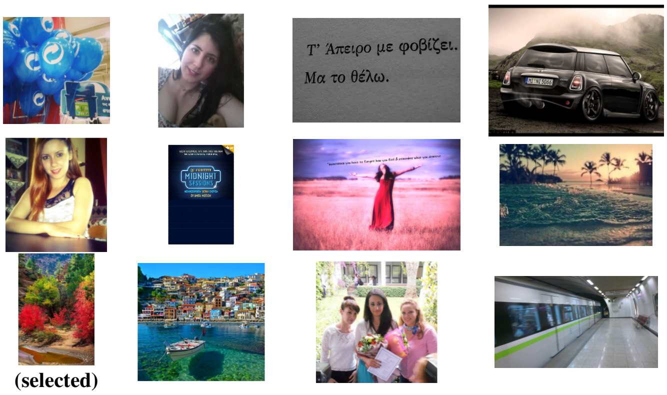


Figure 3: The 12 initial images presented to U_{87} on login and his selection (9th image - forest).

As it can be observed precision-recall between P_1→P_2 and P_2→P_3 changes more rapidly compared to P_3→P_4 and P_4→P_5. This can be explained if we take into consideration that when w_{MCP} increases the retrieved content may not be so visually relevant to the submitted query. On the other hand when w_{MCP} has significantly increased (e.g. > 0.3), then the precision-recall curve seems to converge. This can possibly be explained by the fact that more visually irrelevant content is retrieved, however irrelevance seems to reach a down limit in this dataset (probably a dataset that represents younger ages).

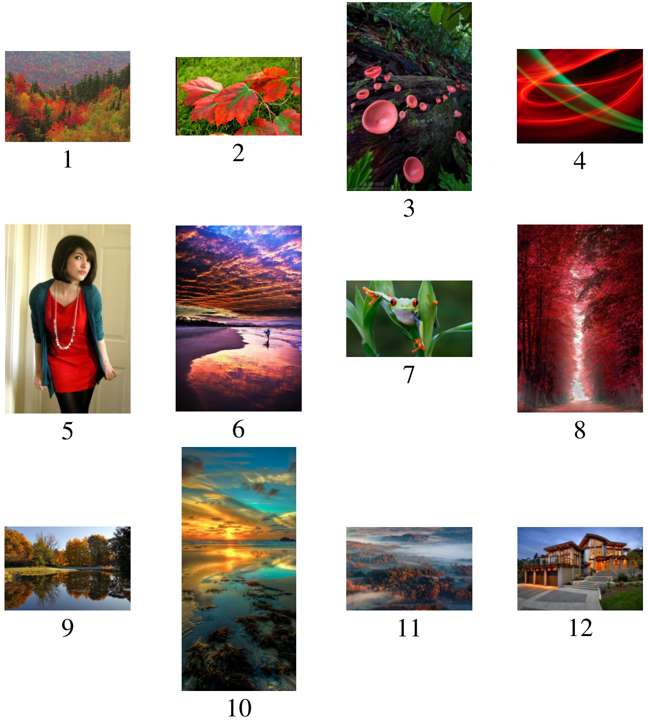


Figure 4: The 12 images retrieved from the database and for profile P_1.

In the following experiment we have also tried to subjectively evaluate the performance of the proposed social relevance feedback scheme. Towards this direction we have used the content visualization tool incorporated in the experiments of [22]. This tool also enables users to select content of interest and iteratively perform queries by example. When users logged in they were presented with the most recent images (based on the time instance the image was posted on Facebook). The tool can present from 4 up to 24 images per instance. In one such experiment user U_{87} has set the limit to 12 images. Furthermore the default setting was the P_1 pair and the user did not change it. In this case 12 images have been presented to the user (Figure 3). Among them, user U_{87} selected the 9th image as relevant to his interests. This image contains a forest and it seemed that this user was interested in other similar images (since he has not changed the P_1 pair). The system responded with 12 new images illustrated in Figure 4, where similar pictures with larger *MCP* are presented first. As it can be observed and since the profile is P1, the retrieved images are Furthermore it should also be mentioned that the retrieval mechanism did not take into consideration the time instance when each image was posted. This means that an image which was posted at the first day of the first month of content gathering, was addressed equally with images posted on the last day of the fourth month.

As it can be observed in Figure 4, since P_1 is used visual relevance is more important than *MCP*. However images with large *MCP* values (e.g. the 4th, 7th or 12th) also appear in the results.

Finally we have also contacted an experiment to test user satisfaction. We have asked 50 different users to evaluate the retrieval performance of P_1, P_2, P_3, P_4 and P_5 by providing a preference. Towards this direction we have recorded 70 sessions for each user. In each session the user provided a query image of her interest and 12 images where retrieved according to P_1, P_2, P_3, P_4 and P_5. Then the user was asked to express preference among the 5 different setups (the question was: which of the 5 setups provides the most satisfactory dozen ?). All results were aggregated and they are provided in Figure 5.

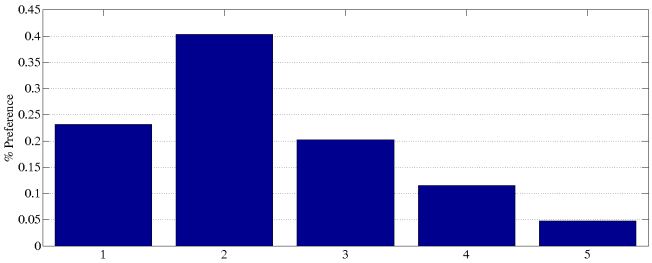


Figure 5: User Preferences for the 5 different setups: 1→P_1, 2→P_2, 3→P_3, 4→P_4, 5→P_5.

P_2 received 40.3% (1,411 sessions), P_1 received 23.2% (812 sessions), P_3 received 20.3% (709 sessions), P_4 received 11.5% (403 sessions) and P_5 reached 4.7% (165 sessions). According to this experiment it seems that users prefer to see content that has attracted the interest of several people. However, when visual similarity is heavily disregarded, the content looks random to users and it is probably far away from specific interests. For example in case of P_5, the retrieval mechanism returns very interesting content, which however may be irrelevant to a user's needs.

VI. Conclusion

CBIR with relevance feedback strategies has the potential to be at the forefront of the technological movement, reducing the pain of learning for a brand new generation of interactive applications. In this framework content modeling and management is an important issue for many new emerging interactive multimedia services. Due however, to the subjective perception of humans as far as the content is concerned, adaptive management algorithms are required to update the system response to actual users' information needs and preferences. However none of the existing approaches meets completely the requirements of an accurate CBIR system with relevance feedback because none of the techniques have completely solved the problem of the semantic gap. So it is still undecided what the future truly holds for improving and implementing RF in real world applications.

Towards improving the performance of RF algorithms, in this paper we have taken into consideration also content power. Our belief is that retrieved content should not only be visually similar but also influential. Since visual relevance by itself cannot provide satisfactory results in terms of semantics. According to this concept, an *MCP* metric has been defined and introduced in the distance between two files. Experiments indicate that the presented social relevance feedback scheme provides promising results.

Future research could include many more experiments with a larger set of users and much more content. Additionally, influential users could also be taken into consideration during the retrieval process. Finally comparison to other typical relevance feedback algorithms could be performed in order to further evaluate user preferences.

Acknowledgment

The authors wish to thank very much all members of the Online Computing Group (www.facebook.com/klimis.ntalianis.7) that have participated in the performed experiments.

References

[1] D. Zeng, C. Hsinchun, R. Lusch, L. Shu-Hsing, “Social Media Analytics and Intelligence,” *IEEE Intelligent Systems*, vol. 25, no. 6, pp. 13–16, 2010.

[2] O. Phelan, K. McCarthy, and B. Smyth, “Using twitter to recommend real-time topical news,” In Proc. of the *3rd ACM conference on Recommender systems*, p.p. 385–388, New York, USA, 2009.

[3] J. Hannon, M. Bennett, and B. Smyth “Recommending twitter users to follow using content and collaborative filtering approaches,” *4th ACM conference on Recommender systems*, p.p. 199–206, New York, USA, 2010.

[4] S. G. Esparza, M. P. O’Mahony, and B. Smyth, “On the real-time web as a source of recommendation knowledge,” In Proc. of the *4th ACM conference on Recommender systems*, p.p. 199–206, New York, USA, 2010.

[5] J.-H. Su, W.-J. Huang, P. S. Yu, and V. S. Tseng, “Efficient Relevance Feedback for Content-Based Image Retrieval by Mining User Navigation Patterns”, *IEEE Transactions on Knowledge and Data Engineering*, 2011

[6] Chueh-Yu Li and Chiou-Ting Hsu, “Image Retrieval with Relevance Feedback Based on Graph- Theoretic Region Correspondence Estimation” *IEEE Trans. Multimedia*, Vol. 10, No. 2, pp. 447-456, April 2008

[7] Wei Bian and Dacheng Tao, “ Biased Discriminate Euclidean Embedding for Content-Based Image Retrieval” *IEEE Trans.on Image Processing*, Vol. 19, No. 2, pp.545-554, Feb 2010.

[8] P. Auer, Z. Hussain, S. Kaski, A. Klami, J. Kujala, J. Laaksonen, A. P. Leung, K. Pasupa and J. Shawe-Taylor, “Pinview: Implicit Feedback in Content-Based Image Retrieval,” *JMLR: Workshop on Applications of Pattern Analysis*, p.p. 51-57, 2010.

[9] D. G. Iowacka and J. Shawe-Taylor, “Content-based Image Retrieval with Multinomial Relevance Feedback,” JMLR: Workshop and Conference Proceedings, *2nd Asian Conference on Machine Learning*, Tokyo, Japan, p.p. 111-125, 2010.

[10] S. Nematipour, J. Shanbehzadeh and R. A. Moghadam, “Relevance Feedback Optimization in Content Based Image Retrieval Via Enhanced Radial Basis Function Network,” Proc. *International Multi-Conference of Engineers and Computer Scientists*, Hong Kong, vol. 1, 2011.

[11] T. Sikora, “The MPEG-7 Visual Standard for Content Description-An Overview,” *IEEE Trans. On CSVT*, Vol. 11, No. 6, pp. 696-702, June 2001

[12] A. Alatan, L. Onural, M. Wollborn, R. Mech, E. Tuncel and T. Sikora, “Image Sequence Analysis for Emerging Interactive Multimedia Services - The European Cost 211 Framework,” *IEEE Trans. Circuits and Systems for Video Tech.*, Vol. 8, No. 7, pp. 802- 813, Nov. 1998.

[13] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int',l J. Computer Vision*, vol. 2, no. 60, pp. 91-110, 2004.

[14] H. Bay, T. Tuytelaars, and L.V. Gool, “Surf: Speeded up robust features,” In *ECCV 2006*, pages 404–417, Springer, 2006.

[15] Y. Avrithis, N. Doulamis, A. Doulamis, and S. Kollias, “Optimization Methods for Key Frames and Scenes Extraction,” *Computer Vision and Image Understanding*, Academic Press, Vol. 75, Nos. 1/2, pp. 3-24, July/August 1999.

[16] N. Doulamis, A. Doulamis, Y. Avrithis, K. Ntalianis, and S. Kollias, "Efficient summarization of stereoscopic video sequences", *IEEE Trans. Circuits Syst. Video Technol.* (Special Issue on 3-D Video Processing), vol. 10, pp.501 -517 2000.

[17] A. D. Doulamis, N. D. Doulamis and S. D. Kollias, "A Fuzzy Video Content Representation for Video Summarization and Content-Based Retrieval," *Signal Processing*, Elsevier Press, Vol. 80, pp. 1049-1067, June 2000.

[18] Y. Ishikawa, R. Subramanya and C. Faloutsos, “Mindreader: Query Databases through Multiple Examples,” Proc. of the *24th VLDB conference*, New York, USA, 1998

[19] E. Kreyszig, *Introductory Functional Analysis with Applications*. New York: Wiley 1989.

[20] A. D. Doulamis and N. Doulamis "Generalized nonlinear relevance feedback for interactive content-based retrieval and organization" *IEEE Trans. Circuits Syst. Video Technol*., vol. 14, no. 5, pp. 656-671, 2004.

[21] S.-H. Lim, S.-W. Kim, S. Park, and J. H. Lee, “Determining Content Power Users in a Blog Network: An Approach and Its Applications,” *IEEE Trans. On Systems, Man and Cybernetics—PART A: Systems and Hymans*, Vol. 41, No. 5, Sept. 2011.

[22] K. Raftopoulos, K. Ntalianis, D. Sourlas and S. Kollias, “Mining User Queries with Markov Chains: Application to Online Image Retrieval,” "Mining User Queries with Markov Chains: Application to Online Image Retrieval.," *IEEE Trans. Knowledge and Data Engineering*, Vol. 25, No. 2, Feb. 2013.

Klimis S. Ntalianis was born in Athens, Greece, in 1975. He received the Diploma degree and the Ph.D degree in electrical and computer engineering, both from the National Technical University of Athens (NTUA), Athens, Greece, in 1998 and 2002 respectively. His Ph.D. studies were supported from the National Scholarships Foundation and the Institute of Communications and Computers Systems of the NTUA. During the last decade Dr. Ntalianis has received six prizes for his academic achievements. He is the author of more than 90 scientific articles, while his research interests include multimedia annotation systems, 3-D image processing, video organization, multimedia security and social computing. He is currently an Assistant Professor at the Technological Educational Institute of Athens.

Anastasios D. Doulamis received the Diploma degree in Electrical and Computer Engineering from the National Technical University of Athens (NTUA) in 1995 with the highest honor. In 2000, he has received the PhD degree in electrical and computer engineering from the NTUA. From 1996-2000, he was with the IVM Lab of the NTUA as research assistant. In 2002, he joined the NTUA as senior researcher. His PhD thesis was supported by the Bodosakis Foundation. In 2006, he was elected Assistant professor at the Technical University of Crete in the area of multimedia systems. Dr. Doulamis has received several awards and prizes during his studies, including the Best Greek Student in the field of engineering in national level in 1995, the Best Graduate Thesis Award in the area of electrical engineering with A. Doulamis in 1996 and several prizes from NTUA, the National Scholarship Foundation and the Technical Chamber of Greece. In 1997, he was given the NTUA Medal as Best Young Engineer. In 2000, he received the best Phd thesis award by the Thomaidion Foundation in conjunction with N. Doulamis. In 2001, he served as technical program chairman of the VLBV'01. He has also served as program committee in several international conferences and workshops. He is reviewer of IEEE journals and conferences as well as and other leading international journals. He is author of more than 200 papers in the above areas, in leading international journals and conferences. His research interests include non-linear analysis, neural networks, multimedia content description, and intelligent techniques for video processing.

Methodology for the modeling of multi-player games

Arturo Yee and Matías Alvarado

***Abstract*—**In this paper a methodology for the algorithmic setting of multi-player games like baseball or American football is presented. We use formal grammars for the whole description language gaming of these sports. Any simple or complex moving play among the team players is obtained according to the rules of the corresponding grammar. The finite state machines for the languages recognition are defined as well. Test validation is by means of sets of hundred simulations that use statistics of real life matches as the input. The output from simulations is highly similar with real life matches results.

***Keywords*—**finite state machine, formal grammar,multi-player games, sport games simulations.

I. Introduction

RECENTLY,the formal modeling and strategic analysis for support the matches gaming of multi-player games, such as, American football (AF) or baseball is growing [4, 13, 20]. These multi-player games have led investigations in areas of sport science [1, 15, 17], computer science, game theory [2], operation research [4, 20], simulation models [11, 13], among others.

The methodology proposed in this paper supports the formal modeling of multi-player games allowinghavingcorrect models of the reality;these models represent the interactions among different entities (players) following a set of rules (game's rules).In addition, the method allowsforstudy behaviors of the different playersregarding the circumstances in a match, and, usingthe statistics of matches being played by human beings,computer simulations can forecast real scores.We emphasize the use of formal grammars and finite state machines for having correct modeling of the games gaming;as well, generators of plays based on statistics to estimatereal scores. This methodology canbe relevant applied in the correct modeling of topics that involve multi-agent contribution.

The automata theory and formal languages theory provide the basis for the algorithms and allow modeling and designs of solutions for huge number of problems. In computer science, one of the main contributions of the study of formal languages is its contribution to the design of programming languages.

The computational automation of multi-player games is supported by finite state machines (FSM) and formal languages. Using FSMs and the formal languages ensure the correct model mathematically, and the correct implementation of algorithms for computer simulation of multi-player games. For algorithmic automations, the formal grammars translate the rules of the games and generate formal languages which describe any possible move during a match game. These formal languages are read by the corresponding FSMs.

Generators of plays produce strings (games' plays) based on the average probability of each of these plays in games played by humans in professional leagues, such as, Major League Baseball (MLB) and National Football League (NFL).

A formal grammar is a set of alphabet symbols and set of rules that produces a formal language. The rules describe how strings have to form from the alphabet symbols and that strings are correct according to the syntax.

The formal grammar is described as follows:

- Visthe alphabet of terminals and non-terminals.
- $\Sigma \subseteq V$is the set of terminals.
- $V-\Sigma$is the set of non-terminal elements.
- $R \subseteq (V-\Sigma) \times V^*$is the set of rules.
- $B \in V-\Sigma$is the initial symbol.

In multi-player games, the formal languages provide a way for describing the games, i.e., they describe all different ways that a game can be played, since; these languages are generated by formal grammars which translate the games' rules.

A FSM is a mathematical device for reading the input strings from a formal language; the parsing of strings starts at the FSM initial state then following through intermediate states; whenever the output or parsing end of a string occurs in a FSM halt state conclusion is that the string belongs to the language being recognized by this FSM [5, 14]. A FSM reads the formal language of the game.

A formal description of FSM is as follows. Let $(\Sigma, \hat{S}, s_0, \varphi, H)$be FSM such that.

- $\Sigma$is the alphabet.
- $\hat{S} = \{s_0, \dots, s_n\}$is the set of states.
- $\varphi: \hat{S} \times \Sigma \rightarrow \hat{S}$is the transitions function.
- $s_0 \in \hat{S}$is the initial state.
- $H = \{s_0, \dots, s_m\} \subseteq \hat{S}$is the set of halt states.

The generator of plays produces strings that must have a

Arturo Yee is Ph. D. students in Computer Science Department at CINVESTAV-IPN. México City, México. Phone +52 55 5747 3756 Ext. 6555; e-mail: ayee@computacion.cs.cinvestav.mx.

Matías Alvarado is a research scientist in Computer Science Department at CINVESTAV-IPN. México City, México.Phone +52 55 5747 3756 Ext. 6555, e-mail: matias@cs.cinvestav.mx.

correct sequence of moves, i.e., the games' plays should generate according to their average frequency of occurrence in real life games and the sequence should be consistent with reality, for this purpose, real statistics of teams who play in MLB and NFL are used.

A generator of plays is useful, because it generates valid baseball strings randomly, quickly and easily. The generator produces games' plays and verifies that.

- These must be made based on their average occurrence frequency, and also
- Be generated following the rules of the game.

Computer simulations team NFL/MLB matches are obtained using instances of our methodology, these simulations are reported in the experimental stage.

The rest of the paper is organized as follows: Section II concerns the baseball modeling. Section III describes American football modeling. Section IV shows the methodology applications in computer simulations. Section 5 presents a discussion and the paper closes with conclusions.

II. Baseball Modeling

Baseball is a multi-player game, played on a diamond-shaped field, two teams confronted during the nine ordinary innings of the match; an inning is complete when both teams have played the offensive and defensive role; the offensive role goal is to score runs while the defensive role is to record 3 outs of the adversary; extra innings are allowed when the match score is tied at the ninth inning. The team that scores more runs at the end of the match is the winner [6, 7].

Baseball analysis has been practiced on diverse issues for this game, including medical and health [21], psychological-emotional [16], specific players' performance [10], or on the best team formation regarding the available players and their skills [8].

In [2], the automation of baseball gaming comprises the basic and compound defense or offence plays by i player; baseball basic plays are weighted and the total is ordered regarding the frequency of their occurrence from MLB (Major League Baseball) statistics, e.g., *strike* occurs more frequently than *hit*, being precision weighted from our own computer simulation matches; the formal grammar rules set the generation of any simple or complex baseball gaming description, including a whole match; the baseball formal language is read by the associated finite state machine (FSM), hence any simple or complex baseball expression is formally correct; the occurrence of plays is in a realistic manner such that the higher the frequency of occurrence of a play in real human matches, the higher the probability that the play is included in the formal account and simulation of the match; the FSM for baseball is modeled like a shape-of-field: the home, 1st, 2nd and 3rd bases are modeled as the FSM states.

A. *Formal language*

The grammar rules generate the formal language (set of strings over the alphabet of symbols of simple plays) describing the whole baseball match, rules ensure the correct composition of sentences describing the baseball plays and matches. The baseball grammar is represented in Table I - Table III.

Table I. Terminals symbols to simple plays.

b^i: ball	
bo^i: bolk	co^i: contact of ball
bg^i: base hit	h^i: homerun
bp^i: base on balls	hi^i: hit
d^i: doublet	r^i: stealing base
f^i: foul	s^i: strike
dp^i: double play	t^i: triple
fs^i: sacrifice fly	…

Table II. Non-terminals symbols.

A:	Action by ball contact
B:	Bat
B3:	Bat with three *outs*
M:	Movement
MH:	Home run movement
MR:	Stolen base movement
…	

Table III. Some grammar rules; H is used for hitting abbreviation

B -> b^i B H lead to ball, and hit back
B -> bp^i MG B H generate base on balls, making M and H return (4 balls later)
B ->s^i B H generate a strike and hit back
B -> p^i B H lead punch and hit back (3 strike later)
B -> p^i B3 H lead punch and hit back with three out (3 strikes and 2 outs later)
B -> f^i B H generate a foul, hitting back
B -> d^i MD B H generate a double, moving back to bat
B -> t^i MT B H generate a triplet, M and hit back
A ->hi^i M B Action to generate a hit
A->o^i B Action to generate one out, H back
B -> h^i MH B H a home run generate M and hit back
B ->tb^i M B H generate a bunt, moving and return to bat
B ->tb^i M o^iB H generate a bunt, moving, out and return to bat
B ->tb^i M o^iB3 H generate a bunt, moving, out to bat and team change (2 outs later)
…

B. *Finite state machine*

The FSM that recognizes the formal languages is as follows. Let $(\Sigma, \hat{S}, s_0, \varphi, H)$be a push-down FSM such that.

- $\Sigma$is the alphabet.
- $\hat{S} = \{s, s_0, s_1, s_2, s_3\}$is the set of states.
- $\varphi: \hat{S} \times \Sigma \rightarrow \hat{S}$is the transitions function.
- $s_0 \in \hat{S}$is the initial state.
- $H = \{s, s_0\} \subseteq \hat{S}$is the set of halt states.

The baseball formal language is read by the associated FSM, hence any simple or complex baseball expression is formally correct; the occurrence of plays is in a realistic manner such that the higher the frequency of occurrence of a play in real human matches, the higher the probability that the play is included in the formal account and simulation of the match; the FSM for baseball is modeled like a shape-of-field: the home, 1st, 2nd and 3rd bases are modeled as the FSM states, see Fig.1.

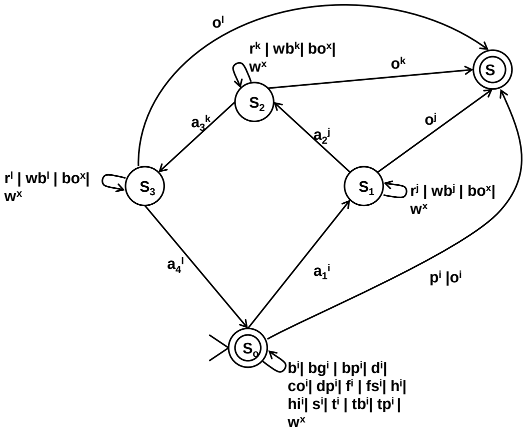


Fig. 1baseball FSM

C. Generator of plays

The generator of baseball plays produces correct strings of sequence of moves by regarding the baseball rules; as well, each baseball play should occurred according to the average frequency of occurrence in real life games.

The generator of strings works once having the baseball play to perform, it has to concatenate with the previous plays at the right end of a string, also indicating the player who performs. The empty string (ε) is for the beginning of a match simulation. Formal grammar, FSM and the generator of random plays are the algorithmic basis for this baseball automation that attains similar scores to human teams' matches in real life, see Fig. 2 and Fig.3.

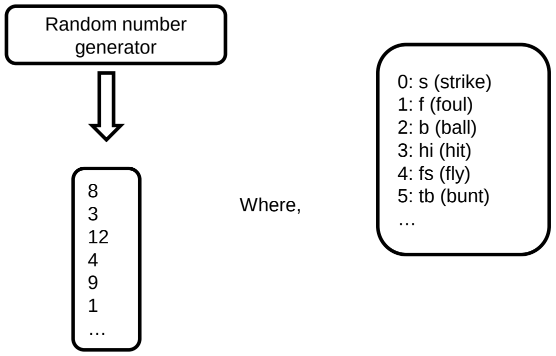


Fig. 2generation scheme of baseball plays

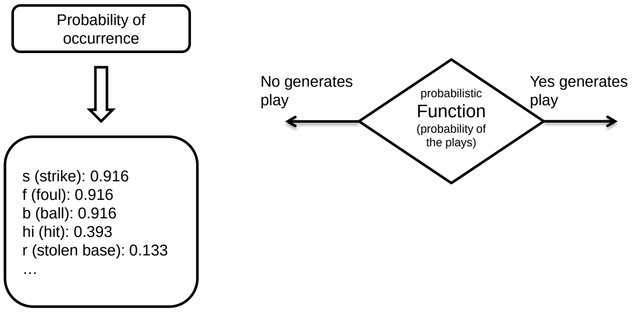


Fig. 3scheme for the use of the probability function

D. String examples

In this section is shown some examples of baseball formal language strings, which are generated by the generator baseball plays and recognized by the baseball finite state described above.

1) String that represents a *homerun* play, player 1 generates a *homerun* batting so the player must move between bases.

$$h^1 a_1^1 a_2^1 a_3^1 a_4^1.$$

In the FSM, this string ends in state s_0, the strings that must pass to be read by s_1, s_2, s_3, and end in state s_0, are those that represent *runs*.

2) String that represents the move by *strikeout*, player 2 at bat, fanning three times consecutive by strikes which causes the player leaves *strikeout*.

$$s^2 s^2 s^2 p^2.$$

In the FSM, this string ends in states$s$, strings that end in s are those that represent *outs*.

3) String that includes many baseball players

$$s^3 s^3 f^3 f^3 co^3 hi^3 a_1^3 s^4 f^4 co^4 hi^4 a_2^3 a_1^4 bg^5 a_3^3 a_2^4$$
$$a_1^5 s^6 co^6 hi^6 a_4^3 a_3^4 a_2^5 a_1^6 s^7 s^7 s^7 p^7.$$

In the FSM, this string passes through all the different states.

III. American football modeling

American football (AF) is one of the top strategic games, played by two teams on a rectangular shaped field, 120 yards long by 53.3 yards wide, with goalposts in the end of the field. Each team has 11 players and a match lasts 1 hour divided in four quarters. The offensive team goal is advance an oval ball, by running or passing toward the adversary's end field [3, 9, 12]. The ways to obtain points are by advancing the ball, ten yards at least, until reach to the end zone for touchdown scoring, or kicking the ball such that it passes in the middle of the adversary's goalposts for a field goal, or by the defensive tackling the ball carrier in the offensive end zone for a safety.

The offensive team should advance the ball at least ten yards in at most four downs (opportunities) to get four additional downs; otherwise the defensive team that is avoiding the ten yards advance, changes to the offensive role. The current offensive team advance starts from the last ball stop position. If the defensive catches the ball before a down is completed, it starts the offensive role at this position. A down ends by the most common circumstances that follow: when a pass is not successful, or when a player is tackled inside the field, or when a ball gets off the field.

A. Formal language

Using the formal grammar rules the generation/description of any simple or complex football gaming, including a whole match, is done. The generated language is read by the associated finite state machine. In order to guarantee the plays occurrence in the simulation, likely to human being matches, we use a generator of random numbers such that the number occurrence carries the occurrence of the associated play.

Let I, I' be different AF teams, $i \in I$ and $i' \in I'$, the formal grammar for American football follows. Let $\hat{G} = (\Sigma, V - \Sigma, R, B)$ be the AF grammar, see Table IV - Table VI.

Table IV. Terminals symbols to simple plays.

kfb^{i}: kick the ball	
cb^{i}: catch the ball	*p*^{i}: punt
rb^{i}: run with the ball	*ga*^{i}: field goal
db^{i}: pass the ball	*re*^{i}: conversion
adb^{i}: advance with the ball	*g*^{i}: goal …
td^{i}: touchdown	

Table V. Non-terminals symbols.

B:	Initial symbol
M:	Movement after kick off
M_1:	Movement for catching the ball
M_2:	Movement for running with the ball
M_3:	Movement for passing the ball
$D_y^{o_i}$:	Denote the downs
M_5:	Auxiliary symbol
M_6:	Auxiliary symbol
M_7:	Auxiliary symbol
…	

Table VI. some grammar rules

$B \rightarrow kfb^{i'}M$ Kick off the ball.

$M \rightarrow \mathrm{cb}^{\mathrm{i}}M_1$The offensive team catches the ball a make a move.

$M_1 \rightarrow rb^iM_2|db^iM_3|\ {}^{\mathrm{j}'}tl^iD_{y=10}^{o_1}$Run, or pass the ball, or the player *i* is tackled by *j*.

$M_2 \rightarrow\ {}^{\mathrm{j}'}tl^iD_{y=10}^{o_1}\Big|td^iT\Big|obD_{y=10}^{o_1}$The player *i* is tackled by *j*, or make a touchdown, or the team is stopped.

$\mathrm{M}_3 \rightarrow cb^iM_1\Big|obD_{y=10}^{o_1}\Big|cb^{j'}M_1'$Catch the ball, or the team is stopped, or interception the ball.

$D_{y=10}^{o_1} \rightarrow D_y^{o_i}$Symbol to define the first down.

$D_y^{o_i} \rightarrow sM_5|p^iM'$Options in the begging of the down.

…

B. Finite state machine

The next finite state machine (FSM) does the algorithmic setting for the AF. Let $(\Sigma, \mathrm{S}, \mathrm{s}_0, \delta, H)$ be a push-down automata such that:

- $\Sigma$is the alphabet.
- $\hat{S} = \{B, M, M_1, M_2, M_3, \dots,\}$is the set of states.
- $\delta\colon S \times \Sigma \rightarrow \hat{S}$is the transitions function.
- $B \in \hat{S}$is the initial state.
- $H = \{M_1', M', B'\} \subseteq \hat{S}$is the set of halt states.

The FSM for the game start, touchdown annotation, and the plays execution in the field are respectively illustrated in Fig. 4, Fig. 5 and Fig. 6.

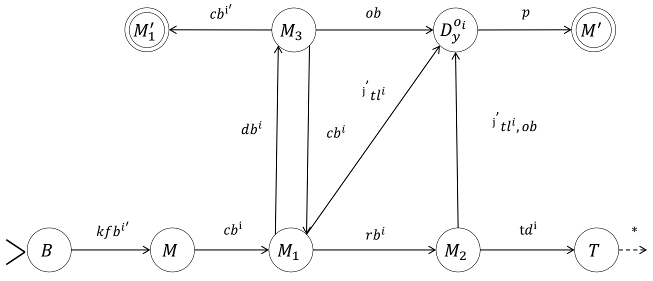


Fig. 4FSM for the game start

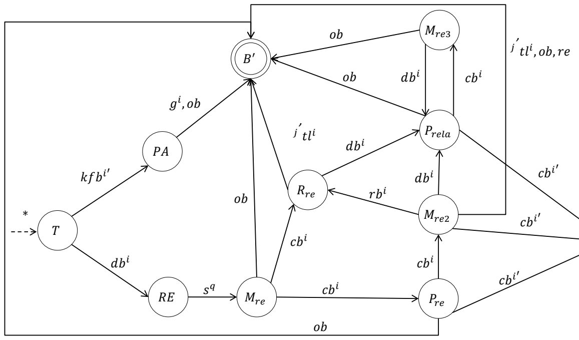


Fig. 5FSM for a touchdown

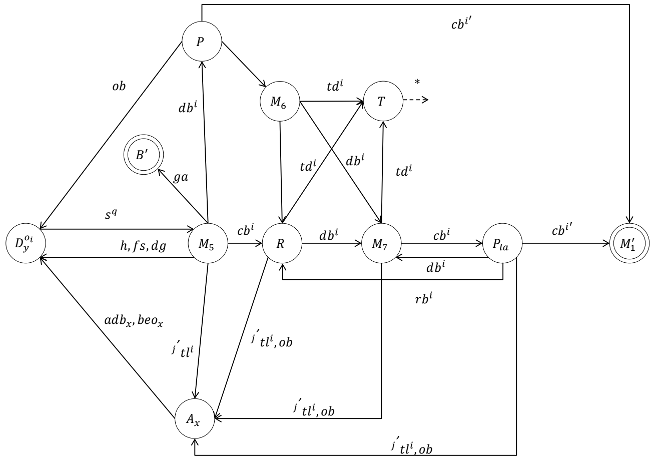


Fig. 6FSM for the plays description in the field

The generator of AF plays as the one of baseball produces correct strings of sequence of moves by regarding the AF rules; as well, each AF play should occurred according to the average frequency of occurrence in real life games.

C. Strings examples

In this section is shown some examples of AF formal language strings.

1) String that represents the beginning of the game, *kickoff* the ball for team 1, the player 7 of the other team *catch* the ball, *run* with ball and he is *tackled* by player 6.

$$kfb^1 team1\ cb^7rb^7tl^6.$$

2) String that represents the *touchdown*, *kickoff* the ball for team 2, the player 9 of the other team *catch* the ball, *run* with ball and makes a *touchdown.*

$$kfb^4 team2\ cb^9rb^9td^9.$$

IV. Methodology Applications in Computer Simulations

In this section, we apply our methodology using statistics of professional teams to do computer simulations.

A. Baseball team performance

To simulate the players' actions according to their performance, we use MLB real statistics from the New York Yankees (NYY) and Oakland Athletics (OAK) in the 2012 season.

Using the MLB statistics [18], the frequency of occurrence of each baseball play per player is used to induce the probability the play can happen in a match.

A total of two hundred computer simulations of baseball matches were performed for the next items, one hundred simulations each.

a. Team 1 (T_1) uses NYY statistics and Team 2 (T_2) uses OAK statistics.

b. T_1 uses OAK statistics and T_2 uses NYY statistics.

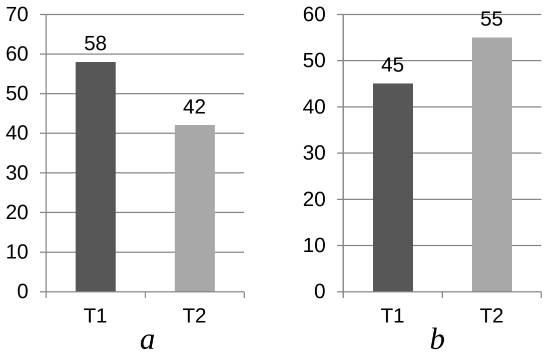


Fig. 7in *a*, 58/42 wins NYY versus OAK one, and in *b*, 45/55 wins OAK team versus NYY one.

In Fig. 7 is shown results when NYY and OAK teams are in competition, the results obtained on these computer simulations are similar ones in season 2012 played by humans.

B. American football team performance

For AF, the NFL statistics from the Denver team and Oakland in the 2012 season were used. The computer simulations results from the data of one hundred AF matches follow. The parameters of the comparison are:

- Point average.
- Victory percent.

The results of computer simulations of one hundred AF matches, Oakland vs. Denver team are reported in TableVII.

Table VII. Results Oakland team vs. Denver team

Parameters	Oakland team	Denver team
Point average	25.53	30
Victory percent	30%	67%

The Denver's scoring, points average per game, is superior to the one of Oakland and therefore won more matches 67/30 percent, respectively. Only the 3 percent of the matches resulted in tie.

V. Discussion

In Game Theory (GT) the formal account of a game models the adversaries' alternate plays to determine the course of actions and strategies during the match of each player and the whole team [19]. The game rules should be unambiguously determined to hold the analysis on competition; different problems have been studied in the perspective of GT such as: modeling, coordination and communication in games and also economics problems such as the gross-domestic product.

The computer simulations of American football and baseball matches are simple and correct. The formal language guaranties the correctness of the computer simulations. The frequency of occurrence of the plays during the human being matches, taken from NFL and MLB statistics, guaranties the real simulation results.

Although baseball and AF are largely different in how to play and the game rules, the formal modeling and the algorithmic setting of both games are similar because the next reason: Baseball and AF are multi-player sport games, but each player has specific roles that do following ordered strategies during the offensive and defensive steps, as one member of a team obeying the manager. Hence, the formal modeling of AF game is quite similar as the one of baseball.

Despite baseball and American football have similarities; there are certain factors in their modeling that must be addressed in a different way.

There is a major factor called *leader play* which is a sequence of moves (plays) made by the players who are playing a direct role in the play.

In baseball, the *leader play* is modeled in such way that all active players' actions are expressed by the formal language in contrast to AF where the *leader play* omits some players' actions, which are not directly part of the play but their contribution affects it, i.e., in AF, different plays in the field are made, although many of them do not do directly on the *leader play* but their the results contributes to the success of *leader play*.

VI. Conclusion

The modeling of multi-player sports gaming is supported by formal languages, finite state machines and statistics-based generators of plays. Hence, computer simulations of matches of baseball and American football arehighly accurate and pretty similar to those performed by human beings. We highlight that this methodology, besides being applied in multi-player games, can be useful to deal with problems where there are interactions among various individuals, each one capable todo particular tasks with specific skills. Besides, this methodology serves as the basis for the use of methods onselection of strategies,necessary to increase the possibilities of a team success.

Acknowledgment

Thanks are extended to the Mexican National Council of Science and Technology (CONACyT) in relation to Arturo Yee's PhD degree grant, CVU 261089.

References

[1] L.W. Alaways, M. Hubbard, Experimental determination of baseball spin and lift, J. Sport Sci., 19 (2001) 349-358.

[2] M. Alvarado, A.Y. Rendón, Nash equilibrium for collective strategic reasoning, Expert Syst. Appl., 39 (2012) 12014-12025.

[3] A.F.C. Association, Offensive Football Strategies, Human Kinetics, Champaign, IL, 2000.

[4] R.D. Baker, I.G. McHale, Forecasting exact scores in National Football League games, Int. J. Forecasting, 29 (2013) 122-130.

[5] D. Barker-Plummer, Turing Machines, The Stanford Encyclopedia of Philosophy, 2005.

[6] P.C. Bjarkman, Diamonds Around the Globe: The Encyclopedia of International Baseball, Greenwood Press, Westport, CT, 2005.

[7] J.C. Bradbury, The Baseball Economist: The Real Game Exposed, The Penguin Group, New York, NY, 2008.

[8] S.S. Britz, M.J.v. Maltitz, Application of the Hungarian Algorithm in Baseball Team Selection and Assignment, in: Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein 2011.

[9] W. Camp, C.S. Badgley, American Football, Createspace Independent Pub, 2009.

[10] W.-C. Chen, A. Johnson, The dynamics of performance space of Major League Baseball pitchers 1871–2006, Ann. Oper. Res., 181 (2010) 287-302.

[11] S.J. Deutsch, P.M. Bradburn, A simulation model for American football plays, Appl. Math. Model., 5 (1981) 13-23.

[12] C. Gifford, American Football Tell me about Sport, Evans Publishing, London, U.K., 2009.

[13] A.J. Gonzalez, D.L. Gross, Learning tactics from a sports game-based simulation, Int. J. Comput. Simul., 5 (1995) 127-148.

[14] J.E. Hopcroft, J.D. Ullman, Introduction to Automata Theory, Languages and Computation, Addison-Wesley, Cambridge, 1979.

[15] T. Jinji, S. Sakurai, Y. Hirano, Factors determining the spin axis of a pitched fastball in baseball, J. Sport Sci., 29 (2011) 761-767.

[16] J.D. Kelly, Reason and magic in the country of baseball, The International Journal of the History of Sport, 28 (2011) 2491-2505.

[17] C. MacMahon, J.L. Starkes, Contextual influences on baseball ball-strike decisions in umpires, players, and controls, J. Sport Sci., 26 (2008) 751-760.

[18] MLB, in, 2012.

[19] J.v. Neumann, O. Morgenstern, Theory of Game and Economic Behavior, 2nd ed., Nueva Jersey: Princenton University Press, 1944.

[20] C. Song, B.L. Boulier, H.O. Stekler, The comparative accuracy of judgmental and model forecasts of American football games, Int. J. Forecasting, 23 (2007) 405-413.

[21] S.J. Thomas, C.B. Swanik, J.S. Higginson, T.W. Kaminski, K.A. Swanik, J.D. Kelly Iv, L.N. Nazarian, Neuromuscular and stiffness adaptations in division I collegiate baseball players, J. oElectromyogr. Kines., 23 (2013) 102-109.

Arturo Yee Rendón got his M. Sc. degree from the Computer Science Department in the Center of Research and Advanced Studies (CINVESTAV-IPN), Mexico, where nowadays is researching on his Ph. D. thesis focus on strategic reasoning for games playing. He got his bachelor degree in Computer Science from the Autonomous University of Sinaloa, Mexico, being distinguished with the Best Student Award in his career promotion.

Control Architecture to Provide E2E security in Interconnected Systems: the (new) SHIELD Approach

Andrea Fiaschetti[1], Andrea Morgagni[2], Andrea Lanna[1], Martina Panfili[1],
Silvano Mignanti[1], Roberto Cusani[3], Gaetano Scarano[3], Antonio Pietrabissa[1],
Vincenzo Suraci[4], Francesco Delli Priscoli[1]

Abstract**—**Modern Systems are usually obtained as incremental composition of proper (smaller and SMART) subsystems interacting through communication interfaces. Such flexible architecture allows the pervasive provisioning of a wide class of services, ranging from multimedia contents delivery, through monitoring data collection, to command and control functionalities. All these services requires that the adequate level of robustness and security is assured at End-to-End (E2E) level, according to user requirements that may vary depending on the specific context or the involved technologies. A flexible methodology to dynamically control the security level of the service being offered is then needed. In this perspective, the authors propose an innovative control architecture able to assure E2E security potentially in any application, by dynamically adapting to the underlying systems and using its resources to "build the security". In particular, the main novelties of this solution are: i) the possibility of dynamically discovering and composing the available functionalities offered by the environment to satisfy the security needs and ii) the possibility of modelling and measuring the security through innovative technology-independent metrics. The results presented in this paper moves from the solutions identified in the pSHIELD project and enrich them with the innovative advances achieved through the nSHIELD research, still ongoing. Both projects have been funded by ARTEMIS-JU.

Keywords**—**Dynamic Composability, E2E Security, Common Criteria, Attack surface metrics, Optimization

I. Introduction

TECHNOLOGICAL advances in computational capabilities along with improvement in communication technologies have enriched the market with a new class of SMART devices that can be used in every application domain, ranging from entertainment, trough automotive and manufacturing, to energy or health.

These devices (i.e. sensor nodes, SMART actuators, programmable controllers, small computing platform, etc.) are commonly referred to as Embedded Devices or Embedded Systems (ESs) and their peculiarities are: i) a reduced size, ii) the possibility of implementing specific functionalities with limited resources and iii) the possibility of interconnecting with other devices to create more complex systems.
Leveraging these peculiarities, several industrial domains have started to massively deploy ESs networks to realize a plenty of tasks, no longer limited to a specific functionality but extended up to end-to-end behaviors.

In order to drive the European research towards an improvement of ES technologies, the European Commission, within the Seventh Framework Programme (FP7) has established the ARTEMIS-JU, a technological initiative in charge of defining and promoting a specific roadmap towards clear and focused objectives [1]. One of these objectives is the development of new technologies and/or strategies to address E2E Security in the context of ESs, with particular care to:

- Solutions oriented to systems certification,
- Cost reduction
- Re-use and re-engineering of non-recurring solutions.

In this context the authors, starting from their academic and industrial backgrounds, have conceived the SHIELD Framework ([2]), an architectural paradigm and design methodology able to address security aspects potentially in each and every domain where ESs (or networks of interconnected ESs) are deployed to provide specific services.
As it happens for communication networks, where modular and cognitive architecture are adopted to provide flexible E2E services that dynamically satisfy the desired level of QoS (see [6]), similarly the interconnection of ESs may require the adoption of a modular and cognitive approach to provide E2E security functionalities that dynamically satisfy the desired "security level" form the end-user.

Thus, the main novelty of the presented approach is the possibility of realizing a "known and predictable" E2E security behavior starting from the composition of individual, atomic elements. In spite of this, the main features of the proposed SHIELD framework are:

1 Authors are with the Department of Computer, Control and Management Engineering "A. Ruberti" at "Sapienza" University of Rome, Via Ariosto 25, 00185 Rome, Italy (e-mail: surname@diag.uniroma1.it).

2 Author is with Selex Electronic Systems (Finmeccanica Company), Via Laurentina 760, 00143 Rome, Italy (e-mail andrea.morgagni@selex-es.com)

3 Authors are with the Department of Information, Communication and Electronic Engineering (DIET) at "Sapienza" University of Rome, Via Eudossiana 18, 00184 Rome, Italy (e-mail: name.surname@ uniroma1.it).

4 Author is with Università degli studi e-Campus, Via Isimbardi 10, Novedrate, 22060, Italy (e-mail vincenzo.suraci@uniecampus.it)

- **modularity** and **expandability** (i.e. the possibility of composing elements together),
- **cognitiveness** and **flexibility** (i.e. the possibility of dynamically adapting to the specific context)
- **technology independence** (i.e. the possibility of abstracting the controlled components in order to measure and provide security in any environment).

The basic approach has already been presented in [2] as preliminary result of the pSHIELD research project ([11]); in this paper an improvement with respect to the basic approach is shown, mainly basing on the recent advances achieved in the execution of the nSHIELD project ([12]), which represents the second phase of the SHIELD Roadmap.

In order to describe the SHIELD approach to E2E security, the rest of the paper is structured as follows: in Section 2 the SHIELD methodology (as presented in [2]) is recalled and in Section 3 the SHIELD behavior as a closed-loop control system is depicted in detail. In Section 4 the innovative control approach to assure E2E security is then presented, and in Section 5 an example is provided. Finally in Section 6 conclusions are drawn.

II. The SHIELD Methodology

The main purpose of the SHIELD methodology is to provide an architectural solution and a design paradigm to enable the Composability of atomic (Security) functionalities in Complex Systems.

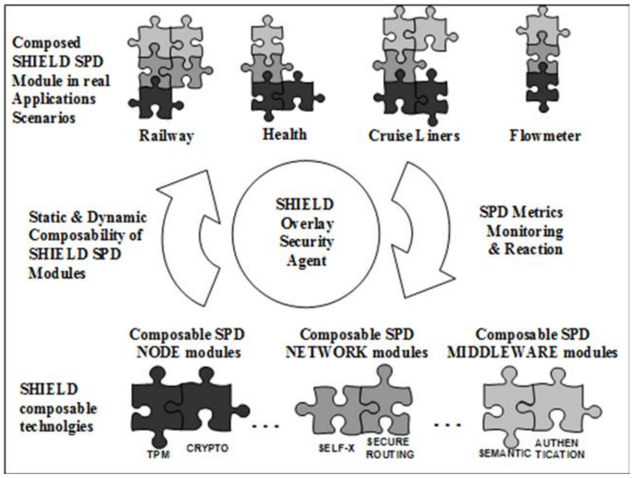

Composed SHIELD SPD Module in real Applications Scenarios
Railway
Health
Cruise Liners
Flowmeter
Static & Dynamic Composability of SHIELD SPD Modules
SHIELD Overlay Security Agent
SPD Metrics Monitoring & Reaction
Composable SPD NODE modules
Composable SPD NETWORK modules
Composable SPD MIDDLEWARE modules
SHIELD composable technolgies
TPM
CRYPTO
SELF-X
SECURE ROUTING
SEMANTIC
AUTHEN TICATION

Fig. 1 SHIELD Methodology

A trivial representation is provided in Fig. 1. The *SHIELD modules* can be represented as pieces of a puzzle, which perfectly fits each other thanks to common interfaces. Each module implements a Security technology or a specific Security functionality. As an example, in Fig. 1 at *node level* there are two modules: Trusted Platform Module and Crypto Technology, at *network level* there are two functionalities: self-x algorithms and secure routing, and at *middleware level* there are two other services: semantic management and authentication.

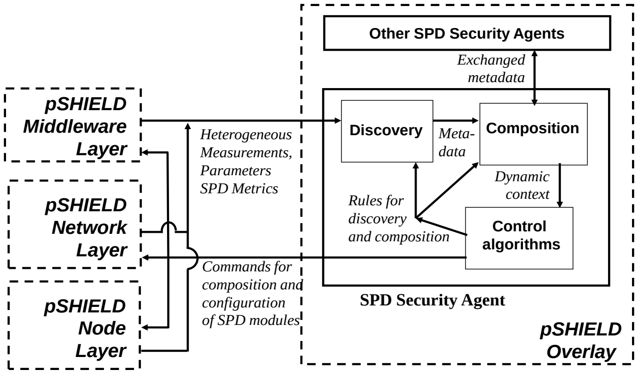


Fig. 2 SHIELD Architecture

These modules, belonging to different SPD layers (node, network or middleware), can be composed (i.e. activated, deactivated or configured) statically or dynamically by the *SHIELD Security Agent*, an innovative software agent (see [4] for details) that collects the information on the system and takes decisions according to proper control algorithms.

This is possible thanks to the development of proper *semantic models* (as outlined in [3] and [5]) that allows the system description in a technology independent way (i.e. machine readable) as well as the definition of *security metrics* that allow the quantification of the security level and consequently the

Thanks to the continuous monitoring performed by the Security Agent, individual SHIELD modules can be dynamically activated and reconfigured once the measured Security metrics do not satisfy the required Security levels, even at run-time.

In addition modularity and technology-independence of the architecture allow a plug&play like behavior, suitable for any kind of application.

In a more structured representation, in Fig. 2 the SHIELD reference architecture is depicted as a control scheme, with the indication of the actors involved in the measurements and commands exchange. The scheme is generically referred to as SPD functionalities, that means Security Privacy and Dependability, since the proposed approach allows to jointly address these peculiarities. However in the prosecution of the paper we will refer only to the “Security” aspects, for which the new metrics and the control algorithms are tailored.

The core of the system, as previously introduced, is the Security Agent: each Agent monitors a set of properly selected measurements and parameters taken from the system (see the arrows labeled as *measurements* in Fig.2). These heterogeneous measurements and parameters are converted by the security agents in *homogeneous/technology-independent) metadata* by extensively using properly selected semantic technologies; the use of homogeneous metadata makes easy the metadata exchange among different security agent (see Fig.2). Each Security Agent, thanks to metadata homogeneity, can aggregate the available metadata, in order to deduce

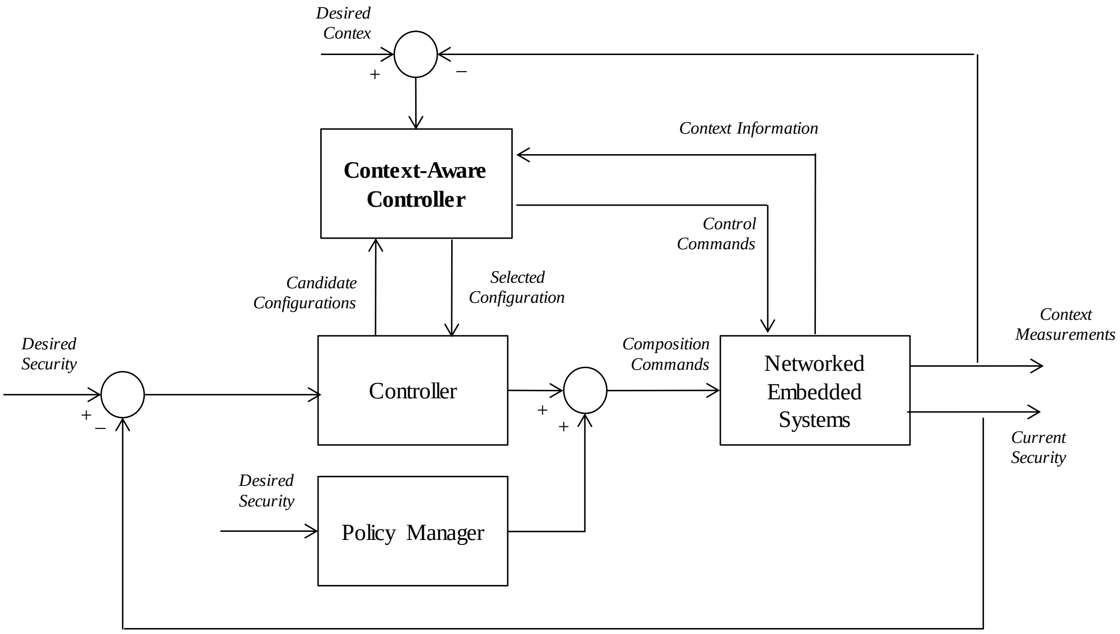


Fig. 3 Composability: a closed-loop view

information which form the so-called *dynamic context* on which the control decisions will be tailored.

Last, but not least, in the security agent runs a set of *control algorithms* which are responsible of dynamically deciding which Security modules have to be composed (i.e. enabled/ disabled/configured) in order to achieve the desired Security level. The decision is driven by the computation of proper technology independent *metrics*, specifically designed for security applications.

In the scope of end-to-end security (the focus subject of this paper), the strength of the SHIELD methodology is that is possible to derive an overall (or E2E) behaviour starting from the atomic behaviors of atomic components and adequately composing them according to proper rules and control algorithms.

On a practical point of view, the SHIELD paradigm allows to deploy small devices (or to use the ones already available), interconnect them and, with the introduction of an intelligent software Agent, dynamically organizing and structuring them so that their capabilities are leveraged to jointly produce the desired effect. As an example, one may be interested in realising the secure monitoring of a train station:

IF the devices deployed in the station (i.e. sensors, cameras, controllers, actuators, etc.) are SHIELD compliant

AND IF at least one SHIELD Security Agent is introduced in this system

THEN it is possible to activate the automatic composition and the system will automatically discover the available devices and the context information, quantify the security level according to the defined metrics, compute a control action and enforce it in the systems to activate/configure the sensors and cameras in the station so that the collected monitoring data are cyphered and made available only to authorized personnel.

This is a trivial example, but is representative of what we call E2E security behaviour: each component is in charge of a specific functionality that is useful to reach the overall objective.

III. THE SHIELD CLOSED-LOOP CONTROL APPROACH

The problem of composing security functionalities can be successfully modelled by leveraging a control theoretic approach (see Fig. 3). Indeed, such kind of model is by far closer to the effective implementation of the SHIELD system.

The *reference signal* is the desired security level, obtained and quantified according to the SHIELD metrics (that will be presented in the following section).

This signal is then used by the *Controller*, that is able to elaborate decision according to proper control algorithms as well as through the interaction with a secondary *Context Controller* that translates ancillary information on the system into constraints and parameters relevant for security purposes. A secondary reference signal may be applied to the system, if, apart from the E2E security behavior, it is also of importance to control other parameters not relevant for security.

In [2] a control algorithm based on Common Criteria composition engine enriched with Hybrid Automata and Model Predictive Control optimization have been proposed as preliminary instantiation of such architecture. This approach has been conceived to be fully in line with the concepts being developed in similar context (e.g. the Future Internet framework [3]) where the limitations coming from the lack of coordination among elements belonging to different layers and/or heterogeneous environments, are addressed through the design of modular controllers and multi-objective procedures.

This solution proved to be valid, but less effective for complex implementations mainly due to the effort needed to translate the “information” into semantic models. The nSHIELD research has then lead to the definition of a new, simpler and more efficient approach, based on these pillars:

- A new metric has been introduced, based on the concept of “attack-surface”, that enables an ease abstraction with respect to the underlying technologies.
- The Common Criteria (CC) guidelines have been confirmed, since the satisfaction of security properties must base its foundations on a consolidated standard, and embedded in the new metrics
- The control algorithm has become the translation of the metrics into an optimization problem, whose objective is to find the elements that maximize a target function

IV. Innovative Control approach to E2E security

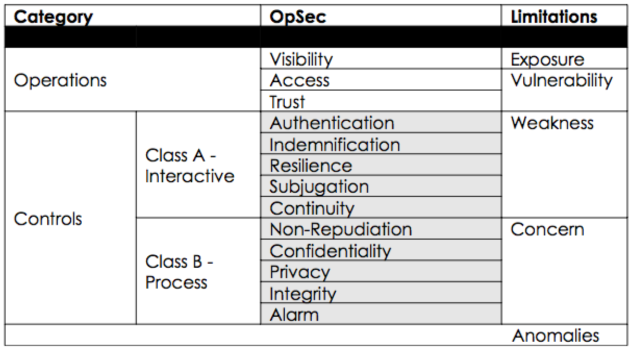
The main novelty of this approach is the definition of “attack surface”, i.e. a virtual line that surrounds a system and by which is possible to identify the potential menaces or vulnerabilities that affect the security level. When composing two or more elements, the attack surface is updated and the new menaces/vulnerabilities are updated as well. Once the final shape of the system is achieved, the resulting surface is the starting point to perform control. As anticipated before, the innovative approach to compose atomic functionalities to achieve E2E security, is based on the two most important standards in cyber-security: Common Criteria (CC, [8]) and Open Source Security Testing Methodology Manual (OSSTMM, [9]).

The OSSTMM “is a methodology to test the operational security of physical locations, human interactions, and all forms of communications such as wireless, wired, analog, and digital”[9]. It is based on the concept of *control* that is the mean to influence the impact of threats and their effects when interaction is required. Controls are divided in two categories:

- **Interactive Controls**, which are able to directly influence visibility, access, or trust interactions and this set includes Authentication, Indemnification, Resilience, Subjugation and Continuity
- **Process Controls**, which do not influence the interactions but they are used to create the defensive processes. They are Non-repudiation, Confidentially, Privacy, Integrity and Alarm.

The activation of a single or a multiple control may originate undesirable effects on the attack surface of the system (i.e. the set of interfaces and vulnerabilities that affect the . The OSSTMM models this element through the *Limitation* concept, which denotes the inability of a control to protect a part of the system. The Limitation value is given by the capabilities of the system and the controls in terms of Vulnerability, Weakness, Concern, Exposure and Anomaly. Fig. 4 shows how the Limitations are mapped with respect to the system and the controls.

In the nSHIELD approach hereby presented, we have improved the OSSTMM standard by considering an attack surface described through the Common Criteria. In particular it has been defined the **Damage Effort Ratio** (DER) as the ratio between the “Damage Potential” and the “Effort” values for each interface, thus obtaining a numerical indicator of the damage that can be caused to the system if a malicious access occurs in this interface. This is a way to measure the “surface” without a-priori knowledge about the system.



Category		OpSec	Limitations
Operations		Visibility	Exposure
		Access	Vulnerability
		Trust	
Controls	Class A - Interactive	Authentication	Weakness
		Indemnification	
		Resilience	
		Subjugation	
		Continuity	
	Class B - Process	Non-Repudiation	Concern
		Confidentiality	
		Privacy	
		Integrity	
		Alarm	
			Anomalies

Fig. 4 Limitations effects

As an example, we could consider an interface in which it is possible to access with three different privileges and three different access rights as in Tab. 1: an interface with “root” privilege and “admin” right has DER=1, WHILE an “authenticated”-“authenticated” combination assures a DER=0,67.

Method Privilege	Value	Access Rights	Value
root	4	admin	4
debugger	3	authenticated	3
authenticated	2	anonymous	2

Tab 1.Example of DER

Considering the inclusion of CC in the OSSTMM standard, the control scheme presented in Fig. 3 is instantiated as depicted in Figure 5:

- the main controller is based on an optimization function that tries to minimize the vulnerability of the attack surface by activating functionalities
- The Context Aware controller has become the OSSTMM controller, since it uses context information to provide the list of Interactive/Process Controls that the main controller may put in place to cope with the security needs.

In addition, the influence of Policy Management (that in [2] has been modelled as a disturb) has become a “controllable” input for the context controller, that considers Policies as constraints to the Interactive/Process controls that it can put in place.

.

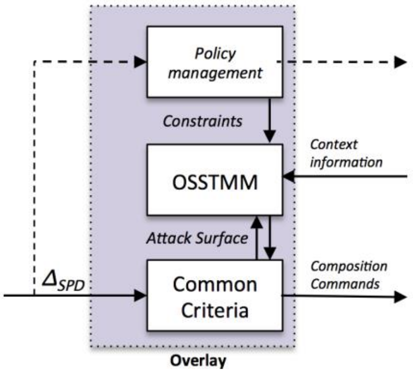


Fig. 5 Innovative Controller

From the mathematical point of view, the main controller solves a typical optimal control problem where the objective function is the minimization of Security value and the constraints is given by the OSSTMM-CC standards and by the policy management system. Higher values of Security cause the activation of more controls and countermeasures; this is the reason why the optimal problem minimizes the Security value. In particular, it minimizes the Δ_{SPD} (i.e. security) value, which is the difference between the desired and actual SPD values.

V. Example of the new SHIELD approach

The example by which the proposed methodology has been tested is an improvement of the one presented in [2] as final demonstration of the pSHIELD project, i.e. the “Monitoring of freight trains transporting hazardous material”.

The hypothesized platform is composed by a central unit connected by means of a ciphered wireless network to remote sensors. In this platform the assets to protect are data sent by remote sensors to central unit, where data are recorded inside the central unit itself.
Threats identified for the above scenario are the following:

- Unauthorized disclosure of information stored within or communicated through computers or communications systems;
- Unauthorized modification or destruction of stored information;
- Manipulation of computer or telecommunications services resulting in various violations;
- Propagation of false or misleading information;
- Users lacking guidance or security awareness;
- Data entry or utilization error;
- Faulty access rights management;

Security functionalities (i.e. Controls) that counter the above threats belong to the following categories:

- Authentication;
- Confidentiality;
- Non repudiation;
- Subjugation.

The application of the surface Attach metrics approach does not depend on a thorough knowledge of the theory that generates such an approach, but only by a well-established knowledge that the supplier of the system and/or components of a system must have on security issues.

Starting from the previously evidenced threats, for each of the two components the attack surface value must be computed, according to the guidelines provided in [8] and [9].

The values for the components of the sample scenario are:

- Central unit: 88,75

Constant due to the lack of controls that could be implemented

- Wireless Sensor Network: [84,089 93,340]

Depending on which of the two available controls is activated. In fact it is important to consider that the different choice of key management and Cryptographic operation algorithm change the vulnerability type, so it insert the possibility, changing these algorithm to modify the Security level of the component introducing different states.

In this case the formulation of an Optimization function is not needed, since it is evident that the most robust configuration is the one associated to a 93,340 value for the WSN. However, in case the available controls and their combination is very high, it is sufficient to maximize the Optimization function given by the sum of the atomic security value, within the constraints defined by policies (i.e. mutual inclusion or mutual exclusions of controls).

VI. Conclusions and Future Works

In this paper the innovative results achieved by the nSHIELD project have been presented, as a significant improvement of the proof of concept reported in [2]. In particular it has been shown how it is possible to drive an E2E behavior by acting on the atomic elements; the key idea is to describe each component with a clear and univocally defined metric value that measure the vulnerability of its attack surface (derived as a mix of [8] and [9] guidelines). Then, while composing together several elements, the resulting attack surface is obtained as the result of an optimization problem whose potential solutions are the different controls that the atomic elements can put in place to countermeasure specific menaces. The problem may be solved by exploration or through simple heuristics.

The proposed methodology is currently being intensively tested in industrially relevant scenarios from the avionic and railways domains and the results will be made available in the final nSHIELD project deliverables.

Future works foresee the adaptation of the proposed approach to address also other problems. It could be particularly helpful, for example, in scenarios where the topologies change very often and the E2E behavior is the provisioning of a specific service, like power distribution (see [10]). The main challenge will be the adaptation/tailoring of a proper metric to the new domain, since a good metric is the basis of any SHIELD-like methodology.

ACKNOWLEDGMENT

The authors would like to thank the pSHIELD and nSHIELD partners for the significant work performed in these projects to transfer the SHIELD methodology from idea to reality.

REFERENCES

[1] ARTEMIS Strategic Research Agenda, March 2006

[2] Fiaschetti A., Suraci V., Delli Priscoli F. *"The SHIELD Framework: how to control Security, Privacy and Dependability in Complex Systems"*, Proceedings of IEEE Workshop on Complexity in Engineering (COMPENG2012), June 11-13, Aachen, 2012

[3] Fiaschetti A., Suraci V., Delli Priscoli F., Taglialatela A., *"Semantic technologies to model and control the "composability" of complex systems: a case study"*, Book chapter of 'Semantics: Theory, Logic and Role in Programming', Nova Publisher, 2012.

[4] Suraci V., Fiaschetti A., Anzidei G., *"Design and implementation of a service discovery and composition framework for security, privacy and dependability control"*, Future Network & Mobile Summit 2012, July 2012, Berlin, Germany

[5] Fiaschetti A., Lavorato F., Suraci V., Palo A., Taglialatela A., Morgagni A., Baldelli A., Flammini F., *"On the use of semantic technologies to model and control Security, Privacy and Dependability in complex systems"* Proc. Of 30th International Conference on. Computer Safety, Reliability and Security (SAFECOMP'11), Sep. 2011. Naples, Italy.

[6] Castrucci, M., Delli Priscoli, F., Pietrabissa, A., Suraci, V., *"A cognitive future internet architecture"* (2011) Lecture Notes in Computer Science, 6656, pp. 91-102, ISBN: 978-364220897-3, doi: 10.1007/978-3-642-20898-0_7

[7] F. Delli Priscoli, *"A fully cognitive approach for future internet"*, Future Internet, vol. 2, no. 1, pp. 16–29, 2010.

[8] Common Criteria for Information Technology Security Evaluation, v3.1, July 2009

[9] OSSTMM, Open Source Security Testing Methodology Manual

[10] S. Canale, A. Di Giorgio, A. Lanna, A. Mercurio, M. Panfili, and A. Pietrabissa, *"Optimal planning and routing in medium voltage powerline communications networks"*, IEEE Trans. Smart Grid, vol. 4, no. 2, pp. 711–719, 2013.

[11] pSHIELD Technical Annex, June 2010

[12] nSHIELD Technical Annex, September 2011

Andrea Fiaschetti obtained his Ms.C. degree in Control Systems Engineering in 2009 fron the University of Rome "La Sapienza" and his Ph.D. degree in Systems Engineering in 2013 from the same University. His research interest is in the field of control theory applied to security domains, with particular focus on modular and composable architecture. He is also vice-president of the Complex Systems Engineering Committee, within

Andrea Morgagni obtained his Ms.C. degree in Biomedical Engineering in 1996 from "Università Politecnica delle Marche". He is currently employed in Selex Electronic Systems (a Finmeccanica Company), where he works as a Senior Security Evaluator for the Evaluation Facilities accredited by the National Security Authority and OCSI . His main expertise is in the field of Security Metrics and Security Certification process for industrial (civil/military) products, according to the ITSEC and Common Criteria (ISO/IEC 15408) standard guidelines.

Andrea Lanna was born in Velletri (Italy) in 1985. He received the Laurea Triennale and Laurea Magistrale degrees in Control and Systems Engineering from the University of Rome "Sapienza", Italy, in 2009 and 2011, respectively, where he is currently pursuing the Ph.D. degree in Systems Engineering and Operative Research. Since December 2011 he has also been working with research group at Value Up s.r.l., Rome. His main research interests include critical infrastructures protection and the application of control systems theory for renewable energy integration in transmission and distribution network. He is involved in Italian and European Research Projects.

Martina Panfili was born in Ceccano (Italy) in 1982. She obtained the Master Degree in System Engineering in 2010 and the Ph.D in System Engineering in 2014 at the University of Rome "Sapienza". Her main research activities are focused on the application of system and control theory to network resource management and Security in the Embedded Systems context. She has been involved in EU research project (MONET, DLC+VIT4IP, nShield).

Silvano Mignanti received his PhD in System Engineering from the University of Rome "Sapienza", Italy in March 2009. He is collaborating with Sapienza since 2005, working in different European projects, among wihch DAIDALOS I and II, WEIRD, P2PNext, Bravehealth, nSHIELD, DLC+VIT4P. He is also collaborating with the CRAT and the CRMPA consortia. Since 2012 he is researcher at Value Up s.r.l.; since april 2014 he is working with Selex-ES in the Fidelity project.

Roberto Cusani received the "Laurea" degree in Electronic Engineering and the Ph.D. in Communication Systems and Computer Science from the University of Rome "La Sapienza". From 1986 to 1990 he was research engineer at the University of Rome "Tor Vergata", teaching Digital Signal Processing. In 1991 he joined the University of Rome "Sapienza" as Associate Professor of Signal Theory. In 2000 he becomes Full Professor and teaches Information Theory and Coding, and Mobile Communications. From 2004 to 2009 he is the head of the Telecommunication Department (INFOCOM) of the University of Rome "Sapienza". He is author of more than 100 publications in international journals and conferences, of the text-book "Teoria dei Segnali" and of five patents regarding telecommunication applications. He was involved in many research programs, both national and international, and in projects with the industries.

Gaetano Scarano was born in Campobasso (Italy) in 1956. He graduated in Electronic Engineering in 1982 at University of Rome "Sapienza". Since 1991 he is working at the University of Rome "Sapienza" where, at present, he is Full Professor and holds the courses "Signal Theory" and "Image Processing and Transmission". His main research activities are on formal methods and on theory of the images transmission. He is the author of about 100 papers appeared on major international reviews and conferences and of one patents. He was/is the scientific responsible, at the University of Rome "Sapienza", for several projects financed by the Italian Minister of Education (MIUR). He is Associate Editor for IEEE Signal Processing Letters.

Antonio Pietrabissa graduated in Electronic Engineering from the University of Rome "La Sapienza", in 2000, where he received the Ph.D. in System Engineering in 2004. Since 2010, he is Assistant Professor with the Department of Computer, System and Management Engineering of the University of Rome "La Sapienza". He is member of the Technical Committe of the Consortium for the Research in Automation and Telecommunication (CRAT). Since 2000, he has been participating in more than 10 European Union, ESA and National projects on telecommunications. His research focus is the application of system and control theory methodologies to telecommunication networks, with specific interest to the design of resource management protocols (e.g., connection admission control, congestion control, routing, medium access control) for multimedia broadband satellite systems, wireless networks and next-generation heterogeneous networks. He is author of more than 20 journal papers and more than 40 conference papers on these topics.

Vincenzo Suraci graduated in Computer Engineering with 110/110 cum laude in October 2004 at the University of Rome “Sapienza”. In April 2008 he pursued a Ph.D. in Systems Engineering in the department of Computer Systems Science of University of Rome “Sapienza”. Currently he is researcher at e-Campus and senior project manager at University of Rome “Sapienza”. His main research interest is to develop and to adapt advanced control and operational research theories (reinforcement learning, column generation, hybrid automata, and discrete event systems) to solve challenging and emerging engineering problems in the field of security and dependability.

Francesco Delli Priscoli was born in Rome in 1962. He graduated in Electronic Engineering in 1986 and he received the Ph.D. in system engineering from the University of Rome "La Sapienza" in 1991. From 1986 to 1991 he worked in the "Studies and Experimentation" Department of Telespazio (Rome). Since 1991 he is working at the University of Rome "La Sapienza" where, at present, he is Full Professor and holds the courses "Automatic Controls" and "Control of Communication and Energy Networks". In the framework of his activity, he has mainly researched on resource/service/content management procedures and on cognitive techniques for telecommunication and energy networks, by largely adopting control based methodologies. He is the author of about 180 papers appeared on major international reviews (about 65), on books (about 10) and conferences (about 120) and of five patents. He was/is the scientific responsible, at the University of Rome "La Sapienza", for 31 projects financed by the EU and by the European Space Agency (ESA), as well as for many national projects and co-operations with major industries. His present research interests concern closed-loop multi-agent learning techniques for Quality of Experience (QoE) evaluation and QoE assurance in advanced communication and energy networks, as well as all related networking algorithms.

Fast Information Retrieval from Big Data by using Cross Correlation in the Frequency Domain

Hazem M. El-Bakry, Nikos E. Mastorakis, Michael E. Fafalios

***Abstract*—**The objective of storing data is to retrieve it as requested in a fast way. In this paper, a new efficient model for fast retrieving of specific information from big data is presented. Fast neural networks are used to find the best matching between words in query and stored big data. The idea is to accelerate the searching operation in a big data. This is done by applying cross correlation between the given query and the big data in the frequency domain rather than time domain. Furthermore, neural networks are used to retrieve information from big data even these data are noised or distorted. The mathematical prove for the acceleration process is analyzed and a formula for the theoretical speed up ratio is given. Simulation results confirm the theoretical considerations.

***Keywords*— *Big data, Cross Correlation, Frequency Domain, Neural networks*.**

I. Introduction

NOWADAYS big data is term that describes the exponential growth and availability of data, both structured and unstructured [1-6].

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data in a single data set. The target moves due to constant improvement in traditional DBMS technology as well as new databases like NoSQL and their ability to handle larger amounts of data. With this difficulty, new platforms of "big data" tools are being developed to handle various aspects of large quantities of data. Examples of Big Data include Big Science, sensor networks, social networks, big social data analysis, Internet documents, Internet search indexing, call detail records, astronomy, atmospheric science, bio

geochemical, biological, and other complex and often interdisciplinary scientific research, military surveillance, forecasting drive times for new home buyers, medical records, photography archives, video archives, and large-scale e-commerce. When dealing with big data, organizations face difficulties in being able to create, manipulate, and manage theses big data. Analyzing big data is a complex problem in business processing because standard tools and procedures are not designed to search and analyze massive datasets [91-103].

The main objective of this paper is to manipulate massive volumes of both structured and unstructured data. Such volumes are so large that it's difficult to process using traditional database and software techniques. The idea of applying cross correlation between input patterns and stored information was applied successfully in many different applications by using neural networks [3-84]. Here, we make use of these previous results to increase the speed of information retrieval from big data. Neural Networks are used to detect the required information from big data even these data are noised or distorted.

The rest of this paper is organized as follows. The theory of fast information retrieval in the frequency domain is presented in section II. Simulation Results are given in section III. Finally conclusions are given.

II. Fast Information Retrieval by using Cross Correlation Implemented in the Frequency Domain

Information processing by using neural networks is divided into two parts. First neural networks are trained to recognize the input patterns. In the test phase, each position in the incoming matrix is processed and tested for the required data (code) by using neural networks. At each position in the input one dimensional matrix, each sub-matrix is multiplied by a window of weights, which has the same size as the sub-matrix. The outputs of neurons in the hidden layer are multiplied by the weights of the output layer. Thus, we may conclude that the whole problem is a cross correlation between the incoming serial data and the weights of neurons in the hidden layer. The convolution theorem in mathematical analysis says that a

H. M. El-Bakry is with Dept. of Information Systems - Faculty of Computer Science and Information Systems – Mansoura University – Egypt. (phone: +2-050-2349340, fax: +2-050-2221442, e-mail: helbakry5@yahoo.com).
N. E. Mastorakis is with the Hellenic Naval Academy (ASEI), Piraeus, Sector of Electrical Engineering and Computers, Piraeus, Greece
mastor@hna.gr (phone 0030 210 4581370)
M. E. Fafalios is with the Hellenic Naval Academy (ASEI), Piraeus, Greece Sector of Electronics and Communications, Piraeus, Greece
fafalios@hna.gr (phone 0030 210 4581644, Fax 0030210 4181768)

.

convolution of f with h is identical to the result of the following steps: let F and H be the results of the Fourier Transformation of f and h in the frequency domain. Multiply F and H* in the frequency domain point by point and then transform this product into the spatial domain via the inverse Fourier Transform. As a result, these cross correlations can be represented by a product in the frequency domain. Thus, by using cross correlation in the frequency domain, speed up in an order of magnitude can be achieved during the test phase [3-84]. Assume that the size of the input pattern is 1xn. In the test phase, a sub matrix I of size 1xn (sliding window) is extracted from the tested matrix, which has a size of 1xN. Such sub matrix, which contains the input pattern, is fed to the neural network. Let W_i be the matrix of weights between the input sub-matrix and the hidden layer. This vector has a size of 1xn and can be represented as 1xn matrix. The output of hidden neurons h(i) can be calculated as follows [3-83]:

$$h_i = g\left(\sum_{k=1}^{n} W_i(k) I(k) + b_i \right) \tag{1}$$

where g is the activation function and b(i) is the bias of each hidden neuron (i). Equation 1 represents the output of each hidden neuron for a particular sub-matrix I. It can be obtained to the whole input matrix Z as follows [2]:

$$h_i(u) = g\left(\sum_{k=-n/2}^{n/2} W_i(k)\ Z(u\ + k)\ + b_i \right) \tag{2}$$

Eq.1 represents a cross correlation operation. Given any two functions f and d, their cross correlation can be obtained by [104]:

$$d(x) \otimes f(x) = \left(\sum_{n=-\infty}^{\infty} f(x+n) d(n) \right) \tag{3}$$

Therefore, Eq. 2 may be written as follows [3-83]:

$$h_i = g\left(W_i \otimes Z + b_i \right) \tag{4}$$

where h_i is the output of the hidden neuron (i) and h_i (u) is the activity of the hidden unit (i) when the sliding window is located at position (u) and (u) $\in$ [N-n+1].

Now, the above cross correlation can be expressed in terms of one dimensional Fast Fourier Transform as follows [3-83]:

$$W_i \otimes Z = F^{-1}\left(F(Z) \bullet F^*\left(W_i \right) \right) \tag{5}$$

Hence, by evaluating this cross correlation, a speed up ratio can be obtained comparable to traditional neural networks. Also, the final output of the neural network can be evaluated as follows:

$$O(u) = g\left(\sum_{i=1}^{q} W_o(i)\, h_i(u) + b_o \right) \tag{6}$$

where q is the number of neurons in the hidden layer. O(u) is the output of the neural network when the sliding window located at the position (u) in the input matrix Z. W_o is the weight matrix between hidden and output layer.

The complexity of cross correlation in the frequency domain can be analyzed as follows:

1- For a tested matrix of 1xN elements, the 1D-FFT requires a number equal to $N\log_2 N$ of complex computation steps [2]. Also, the same number of complex computation steps is required for computing the 1D-FFT of the weight matrix at each neuron in the hidden layer.

2- At each neuron in the hidden layer, the inverse 1D-FFT is computed. Therefore, q backward and (1+q) forward transforms have to be computed. Therefore, for a given matrix under test, the total number of operations required to compute the 1D-FFT is (2q+1) $N\log_2 N$.

3- The number of computation steps required by fast neural networks (FNNs) is complex and must be converted into a real version. It is known that, the one dimensional Fast Fourier Transform requires $(N/2)\log_2 N$ complex multiplications and $N\log_2 N$ complex additions [105]. Every complex multiplication is realized by six real floating point operations and every complex addition is implemented by two real floating point operations. Therefore, the total number of computation steps required to obtain the 1D-FFT of a 1xN matrix is:

$$\rho=6((N/2)\log_2 N) + 2(N\log_2 N) \tag{7}$$

which may be simplified to:

$$\rho=5N\log_2 N \tag{8}$$

4- Both the input and the weight matrices should be dot multiplied in the frequency domain. Thus, a number of complex computation steps equal to qN should be considered. This means 6qN real operations will be added to the number of computation steps required by FNNs.

5- In order to perform cross correlation in the frequency domain, the weight matrix must be extended to have the same size as the input matrix. So, a number of zeros = (N-n) must be added to the weight matrix. This requires a total real number of computation steps = q(N-n) for all neurons. Moreover, after computing the FFT for the weight matrix, the conjugate of this matrix must be obtained. As a result, a real number of computation steps = qN should be added in order to obtain the conjugate of the weight matrix for all neurons. Also, a number of real computation steps equal to N is required to create butterflies complex numbers ($e^{-jk(2\Pi n/N)}$), where 0<K<L. These

(N/2) complex numbers are multiplied by the elements of the input matrix or by previous complex numbers during the computation of FFT. To create a complex number requires two real floating point operations. Thus, the total number of computation steps required for FNNs becomes:

$$\sigma=(2q+1)(5N\log_2 N)+6qN+q(N-n)+qN+N \qquad (9)$$

which can be reformulated as:

$$\sigma=(2q+1)(5N\log_2 N)+q(8N-n)+N \qquad (10)$$

6- Using sliding window of size 1xn for the same matrix of 1xN pixels, q(2n-1)(N-n+1) computation steps are required when using traditional neural networks (TNNs) to process (n) input data. The theoretical speed up factor η can be evaluated as follows:

$$\eta = \frac{q(2n-1)(N-n+1)}{(2q+1)(5N\log_2 N)+q(8N-n)+N} \qquad (11)$$

III. Simulation Results

TNNs accept serial input data with fixed size (n). Therefore, the number of input neurons equals to (n). Instead of treating (n) inputs, the proposed new approach is to collect all the incoming data together in a long vector (for example 100xn). Then the input data is tested by time delay neural networks as a single pattern with length L (L=100xn). Such a test is performed in the frequency domain as described before.

Eq. 11 is also true for recurrent neural networks. The theoretical speed up ratio for processing short successive (n) data in a long input vector (L) using recurrent neural networks is listed in tables 1, 2, and 3. Also, the practical speed up ratio for manipulating matrices of different sizes (L) and different sized weight matrices (n) using a 2.7 GHz processor and MATLAB is shown in table 4. An interesting point is that the memory capacity is reduced when using FNNs. This is because the number of variables is reduced compared to TNNs.

Another point of interest should be noted. In TNNs, if the whole input data (N) is available, then there is a waiting time for each group of (n) input data so that conventional neural networks can release their output for the previous group of (n) data. In contrast, FNNs can process the total N data directly with zero waiting time. For example, if the total (N) input data is appeared at the input neurons, then:

1- TNNs can process only data of size (n) as the number of input neurons = (n).
2- The first group of (n) data is processed by TNNs.
3- The second group of (n) data must wait for a waiting time = τ, where τ is the response time consumed by TNNs for treating each group of (n) input data.
4- The third group of (n) data must wait for a waiting time = 2τ corresponding to the total waiting time required by TNNs for treating the previous two groups.
5- The fourth (n) data must wait for a waiting time = 3τ.
6- The last group of (n) data must wait for a waiting time = (N-n)τ.

As a result, the wasted waiting time in the case of TNNs is (N-n)τ. In the case of FNNs, there is no waiting time as the whole input data (Z) of length (N) will be processed directly and the time consumed is the only time required by FNNs itself to produce their output.

IV. Conclusion

A new approach for testing big data with neural networks has been presented. The operation of neural networks during the test phase has been accelerated. This has been done by applying cross correlation between the whole input patterns and the input weights of neural networks in the frequency domain rather than time domain. The mathematical proof for the acceleration process has been introduced. A formula for the theoretical speed up ratio has been given. Simulation results have confirmed the theoretical considerations.

References

[1] S. del Rio, V. L pez, J.M. Benitez, F. Herrera, On the use of MapReduce for imbalanced big data using Random Forest, ElSEVIER, (2014).

[2] X.-W. Chen, Big Data Deep Learning: Challenges and Perspectives, IEEE, (2014).

[3] Yingyi Bu, Vinayak Borkar, Michael J. Carey, Joshua Rosen, Neoklis Polyzotis, Tyson Condie, Markus Weimer, Raghu Ramakrishnan, Scaling Datalog for Machine Learning on Big Data, (2012).

[4] A. Cassioli , A. Chiavaioli , C. Manes , M. Sciandrone, An incremental least squares algorithm for large scale linear classification, ElSEVIER, (2012).

[5] C.L. Philip Chen, Chun-Yang Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, ElSEVIER, (2013).

[6] Alicia Fernández a, ÁlvaroGómez a, FedericoLecumberry a,n, ÁlvaroPardo b, Ignacio Ramírez a, Pattern Recognitionin Latin Americainthe “Big Data” Era, ElSEVIER,(2014).

[7] Hazem M. El-Bakry, and Mohamed Hamada, “Fast Diagnosing of Pediatric Respiratory Diseases by using High Speed Neural Networks,” Proc. of IEEE IJCNN 2013, Dallas Tx, USA, August 4-9, 2013, pp. 226-232.

[8] Hazem M. El-Bakry, and Nikos Mastorakis, “A New Fast Neural Model,” Proceedings of the 11th WSEAS international conference on Applied Computer and Applied Computational Science (ACACOS'12), Rovaniemi, Finland, April 18-20, 2012, pp 224-231

[9] Hazem M. El-Bakry, "An Efficient Algorithm for Pattern Detection using Combined Classifiers and Data Fusion," Information Fusion Journal, vol. 11, 2010, pp. 133-148.

[11] Hazem M. El-Bakry, "Fast Virus Detection by using High Speed Time Delay Neural Networks," Journal of Computer Virology, vol. 6, no. 2, 2010, pp. 115-122.

[12] Hazem M. El-Bakry, and Nikos Mastorakis, “An Intelligent Approach for Fast Detection of Biological Viruses in DNA Sequence,” Proc. of 10th WSEAS International Conference on APPLICATIONS of COMPUTER ENGINEERING (ACE '11), Spain, March 24-26, 2011, pp. 237-244.

[13] Hazem M. El-bakry, and Nikos Mastorakis, “A New Approach for Prediction by using Integrated Neural Networks,” Proc. of Int. Conf., Mexico, Jan. 29-31, 2011, pp. 17-28.

[14] Hazem M. El-Bakry and Wael A. Awad, “A New Hybrid Neural Model for Real-Time Prediction Applications,” International Journal of Information Technology and Network Application, vol. 1, no. 1, 2011, pp. 1-12.

[15] Hazem M. El-Bakry, “Fast Karnough Map for Simplification of Complex Boolean Functions,” Proc. of 10th WSEAS International Conference on Applied Computer Science (ACS'10), Japan, October 4-6, 2010, pp. 478-483.

[16] Hazem M. El-bakry, and Nikos Mastorakis, “Prediction of Market Price by using Fast Time Delay Neural Networks,” Proc. of 10th WSEAS Int.

Conf. on Neural Networks (NN'10), Romania, June 13-15, 2010, pp. 230-237.

[17] Hazem M. El-bakry, and Wael A. Awad, “Fast Forecasting of Stock Market Prices by using New High Speed Time Delay Neural Networks,” Waset International Journal of Computer and Information Engineering, vol. 4, no.2., 2010, pp. 138-144.

[18] Hazem M. El-bakry, and Nikos Mastorakis, “Fast Packet Detection by using High Speed Time Delay Neural Networks," Proc. of the 10th WSEAS Int. Conference on Multimedia Systems & Signal Processing, Hangzhou University, China, April 11-13, 2010, pp. 222-227.

[19] Hazem M. El-Bakry, Alaa M. Riad, Ahmed Atwan, Sameh Abd El-Ghany, and Nikos Mastorakis “A New Automated Information Retrieval System by using Intelligent Mobile Agent, " Proc. of Recent Advances in Artificial Intelligence, Koweledge Engineering and Databases, Cambridge, UK, February 20-22, 2010, pp. 339-351.

[20] Hazem M. El-Bakry, "New Fast Principal Component Analysis For Real-Time Face Detection," Machine Graphics & Vision Journal (MG&V), vol. 18, no.4, 2009, pp. 405-426.

[21] Hazem M. El-bakry, and Mohamed Hamada “Fast Time Delay Neural Networks for Detecting DNA Coding Regions,” Springer, Lecture Notes on Artificial Intelligence (LNAI 5711), 2009, pp. 334-342.

[22] Hazem M. El-Bakry, "A New Neural Design for Faster Pattern Detection Using Cross Correlation and Matrix Decomposition," Neural World journal, Neural World Journal, 2009, vol. 19, no. 2, pp. 131-164.

[23] Hazem M. El-Bakry, and Ahmed Atwan, " Improving Quality of Business Networks for Information Systems," The International Journal of Computer, Information, and systems science, and Engineering, issue 3, vol. 3, July 2009, pp. 138-150.

[24] Hazem M. El-Bakry, and Ahmed A. Mohammed, "Optimal Document Archiving and Fast Information Retrieval," The International Journal of Computer science, and Engineering, issue 2, vol. 3, July 2009, pp. 108-121.

[25] Hazem M. El-Bakry, and Nikos Mastorakis, "Fast Word Detection in a Speech Using New High Speed Time Delay Neural Networks," WSEAS Transactions on Information Science and Applications, issue 7, vol. 5, July 2009, pp. 261-270.

[26] Hazem M. El-Bakry, and Nikos Mastorakis, "Fast Information Retrieval from Web Pages," WSEAS Transactions on Information Science and Applications, issue 6, vol. 6, June 2009, pp. 1018-1036.

[27] Hazem M. El-Bakry, and Nikos Mastorakis, "Fast Image Matching on Web Pages," WSEAS Transactions on Signal Processing, issue 4, vol. 5, June 2009, pp. 157-166.

[28] Hazem M. El-Bakry, and Nikos Mastorakis, "Fast Detection of Specific Information in Voice Signal over Internet Protocol," WSEAS Transactions on Communications, issue 5, vol. 8, May 2009, pp. 483-494.

[29] Hazem M. El-bakry, and Nikos Mastorakis “Fast Time Delay Neural Networks for Word Detection in a Video Conference,” Proc. of European Computing and Computational Intelligence International Conference, Tbilisi, Georgia, June 26-28, 2009, pp. 120-129.

[30] Alaa M. Riad, Hazem M. El-bakry, and Nikos Mastorakis, “Fast Harmonic Current / Voltage Prediction by using High Speed Time Delay Neural Networks, " Proc. of WSEAS International Conference on Communication and Information, Athens, Greece, December 29-31, 2009, pp. 245-272.

[31] Hazem M. El-Bakry, and Nikos Mastorakis “Fast Human Motion Tracking by using High Speed Neural Networks " Proc. of 9th WSEAS International Conference on SIGNAL, SPEECH AND IMAGE PROCESSING (SSIP '09), Budapest, Hungry, September 3-5, 2009, pp. 221-240.

[32] Hazem M. El-Bakry, and Nikos Mastorakis “A Fast Computerized Method For Automatic Simplification of Boolean Functions,” Proc. of 9th WSEAS International Conference on SYSTEMS THEORY AND SCIENTIFIC COMPUTATION (ISTASC '09), Moscow, Russia**,** August 26-28, 2009, pp. 99-107.

[33] Hazem M. El-Bakry, and Nikos Mastorakis “Fast Information Processing over Business Networks,” Proc. of 9th WSEAS International Conference on Applied Informatics and Communications (AIC'09), Moscow, Russia**,** August 26-28, 2009, pp.305-324.

[34] Hazem M. El-Bakry, and Nikos Mastorakis “A Fast Searching Protocol for Fully Replicated System,” Proc. of of 13th WSEAS International Conference on Computers, Rodos, Greece, July 22-25, 2009, pp. 588-600.

[35] Hazem M. El-Bakry, and Nikos Mastorakis “An Efficient Electronic Archiving Approach for Office Automation,” Proc. of European Computing and Computational Intelligence International Conference, Tbilisi, Georgia, June 26-28, 2009, pp. 130-144.

[36] Hazem M. El-Bakry, and Nikos Mastorakis “Fast Time Delay Neural Networks for Word Detection in a Video Conference,” Proc. of European Computing and Computational Intelligence International Conference, Tbilisi, Georgia, June 26-28, 2009, pp. 120-129.

[37] Hazem M. El-Bakry, “Fast Record Detection in Large Databases Using New High Speed Time Delay Neural Networks,” Proc. of IEEE IJCNN’09, Atlanta, USA, June 14-19, 2009, pp. 757-763.

[39] Hazem M. El-Bakry, and Nikos Mastorakis “Fast Image Matching on Web Pages,” Proc. of Recent Advances in Applied Mathematics and Computational and Information Sciences, Houston, USA, April 30-May 2, 2009, pp. 470-479.

[40] Hazem M. El-Bakry, and Nikos Mastorakis “Design of Anti-GPS for Reasons of Security,” Proc. of Recent Advances in Applied Mathematics and Computational and Information Sciences, Houston, USA, April 30-May 2, 2009, pp. 480-500.

[41] Hazem M. El-Bakry, and Nikos Mastorakis, "A Modified Hopfield Neural Network for Perfect Calculation of Magnetic Resonance Spectroscopy," WSEAS Transactions on Information Science and Applications, issue 12, vol. 5, December 2008, pp. 1654-1666.

[42] Hazem M. El-Bakry, and Nikos Mastorakis, "A New Fast Forecasting Technique using High Speed Neural Networks," WSEAS Transactions on Signal Processing, issue 10, vol. 4, October 2008, pp. 573-595.

[43] Hazem M. El-Bakry, and Nikos Mastorakis, "A New Technique for Detecting Dental Diseases by using High Speed Artificial Neural Network," WSEAS Transactions on Computers, Issue 12, vol. 7, December 2008, pp. 1977-1987.

[44] Hazem M. El-Bakry, and Nikos Mastorakis, "A Real-Time Intrusion Detection Algorithm for Network Security," WSEAS Transactions on Communications, Issue 12, vol. 7, December 2008, pp. 1222-1234.

[45] Hazem M. El-Bakry, and Nikos Mastorakis, " An Effective Method for Detecting Dental Diseases by using Fast Neural Networks," WSEAS Transactions on Biology and Biomedicine, issue 11, vol. 5, November 2008, pp. 293-301.

[46] Hazem M. El-Bakry, and Nikos Mastorakis, "A Novel Fast Kolmogorov’s Spline Complex Network for Pattern Detection," WSEAS Transactions on Systems, Issue 11, vol. 7, November 2008, pp. 1310-1328.

[47] Hazem M. El-Bakry, "New Faster Normalized Neural Networks for Sub-Matrix Detection using Cross Correlation in the Frequency Domain and Matrix Decomposition, " Applied Soft Computing journal, vol. 8, issue 2, March 2008, pp. 1131-1149.

[48] Hazem M. El-Bakry and Mohamed Hamada, "A New Implementation for High Speed Neural Networks in Frequency Space," Lecture Notes in Artificial Intelligence, Springer, KES 2008, Part I, LNAI 5177, pp. 33-40.

[49] Hazem M. El-Bakry, and Nikos Mastorakis “Fast Virus Detection by using High Speed Time Delay Neural Networks,” Proc. of 10th WSEAS Int. Conf. on NEURAL NETWORKS (NN'09), Prague, Czech Repulic, March 22-25, 2008, pp. 169-183.

[50] Hazem M. El-Bakry, and Nikos Mastorakis “New Efficient Neural Networks for Fast Record Detection in Databases,” Proc. of Recent Advances in Artificial Intelligence, Koweledge Engineering and Databases, Cambridge, UK, February 21-23, 2009, pp. 95-102.

[51] Hazem M. El-Bakry, and Nikos Mastorakis “Fast Detection of Specific Information in Voice Signal over Internet Protocol,” Proc. of 7th WSEAS Int. Conf. on COMPUTATIONAL INTELLIGENCE, MAN-MACHINE SYSTEMS and CYBERNETICS (CIMMACS '08), Cairo, EGYPT, Dec. 29-31, 2008, pp. 125-136.

[52] Hazem M. El-Bakry, and Nikos Mastorakis “Information Retrieval Based on Image Detection on Web Pages,” Proc. of 7th WSEAS Int. Conf. on COMPUTATIONAL INTELLIGENCE, MAN-MACHINE SYSTEMS and CYBERNETICS (CIMMACS '08), Cairo, EGYPT, Dec. 29-31, 2008, pp. 221-228.

[53] Hazem M. El-Bakry, and Nikos Mastorakis “Fast Information Retrieval from Web Pages,” Proc. of 7th WSEAS Int. Conf. on COMPUTATIONAL INTELLIGENCE, MAN-MACHINE SYSTEMS and CYBERNETICS (CIMMACS '08), Cairo, EGYPT, Dec. 29-31, 2008, pp. 229-247.

[54] Hazem M. El-Bakry and Mohamed Hamada, "New Fast Decision Tree Classifier for Identifying Protein Coding Regions," Proc. of ISICA ***2008 Conf., China, Dec. 3-5, 2008, pp. 489-500.***

[55] Hazem M. El-Bakry, and Nikos Mastorakis, " An Effective Method for Detecting Dental Diseases by using Fast Neural Networks, " 8th WSEAS International Conference on SIGNAL, SPEECH AND IMAGE PROCESSING (SSIP '08), Santander, Cantabria, Spain, September 23-25, 2008, pp. 144-152.

[56] Hazem M. El-Bakry, and Nikos Mastorakis, " A New Fast Forecasting Technique using High Speed Neural Networks, " 8th WSEAS International Conference on SIGNAL, SPEECH AND IMAGE PROCESSING (SSIP '08), Santander, Cantabria, Spain, September 23-25, 2008, pp. 116-138.

[57] Hazem M. El-Bakry, and Nikos Mastorakis, " Realization of E-University for Distance Learning, " 8th WSEAS International Conference on DISTANCE LEARNING and WEB ENGINEERING (DIWEB '08), Santander, Cantabria, Spain, September 23-25, 2008, pp. 17-31.

[58] Hazem M. El-Bakry, and Nikos Mastorakis, " A Novel Fast Kolmogorov's Spline Complex Network for Pattern Detection," 8th WSEAS International Conference on SIMULATION, MODELLING and OPTIMIZATION (SMO '08), Santander, Cantabria, Spain, September 23-25, 2008, pp. 261-279.

[59] Hazem M. El-Bakry, and Nikos Mastorakis, " A New Technique for Detecting Dental Diseases by using High Speed Neuro-Computers," European Computing Conf. (ECC '08), Malta, September 11-13, 2008, pp. 432-440.

[60] Hazem M. El-Bakry, and Nikos Mastorakis, " A Modified Hopfield Neural Network for Perfect Calculation of Magnetic Resonance Spectroscopy," 1st WSEAS International Conference on Biomedical Electronics and Biomedical Informatics (BEBI '08), Rhodes, Greece, August 20-22, 2008, pp. 242-254.

[61] Hazem M. El-Bakry, and Nikos Mastorakis, " A Real-Time Intrusion Detection Algorithm for Network Security, " 8st WSEAS International Conference on Applied Informatics and Communications (AIC '08), Rhodes, Greece, August 20-22, 2008, pp. 533-545.

[62] Hazem M. El-Bakry, and Nikos Mastorakis "New Fast Normalized Neural Networks for Pattern Detection," Image and Vision Computing Journal, vol. 25, issue 11, 2007, pp. 1767-1784.

[63] Hazem M. El-Bakry, "New Fast Time Delay Neural Networks Using Cross Correlation Performed in the Frequency Domain," Neurocomputing Journal, vol. 69, October 2006, pp. 2360-2363.

[64] Hazem M. El-Bakry and Nikos Mastorakis, "Fast Code Detection Using High Speed Time Delay Neural Networks," Lecture Notes in Computer Science, Springer, vol. 4493, Part III, May 2007, pp. 764-773.

[65] Hazem M. El-Bakry, "New High Speed Normalized Neural Networks for Fast Pattern Discovery on Web Pages," International Journal of Computer Science and Network Security, vol. 6, No. 2A, February 2006, pp. 142-152.

[66] Hazem M. El-Bakry, and Qiangfu Zhao, "Fast Normalized Neural Processors For Pattern Detection Based on Cross Correlation Implemented in the Frequency Domain," Journal of Research and Practice in Information Technology, Vol. 38, No.2, May 2006, pp. 151-170.

[67] Hazem M. El-Bakry, "New Fast Time Delay Neural Networks Using Cross Correlation Performed in the Frequency Domain," Neurocomputing Journal, vol. 69, October 2006, pp. 2360-2363.

[68] Hazem M. El-Bakry, and Nikos Mastorakis, "A Novel Model of Neural Networks for Fast Data Detection," WSEAS Transactions on Computers, Issue 8, vol. 5, November 2006, pp. 1773-1780.

[69] Hazem M. El-Bakry, and Nikos Mastorakis, "A New Approach for Fast Face Detection," WSEAS Transactions on Information Science and Applications, issue 9, vol. 3, September 2006, pp. 1725-1730.

[70] Hazem M. El-Bakry, and Qiangfu Zhao, “Fast Neural Implementation of PCA for Face Detection,” Proc. of IEEE World Congress on Computational Intelligence, IJCNN’06, Vancouver, BC, Canada, July 16-21, 2006, pp. 1785-1790.

[71] Hazem M. El-Bakry, “A Simple Design for High Speed Normalized Neural Networks Implemented in the Frequency Domain for Pattern Detection,” Proc. of IEEE World Congress on Computational Intelligence, IJCNN’06, Vancouver, BC, Canada, July 16-21, 2006, pp. 2296-2303.

[72] Hazem M. El-Bakry, “Fast Co-operative Modular Neural Processors for Human Face Detection,” Proc. of IEEE World Congress on Computational Intelligence, IJCNN’06, Vancouver, BC, Canada, July 16-21, 2006, pp. 2304-2311.

[73] Hazem M. El-Bakry, “New Fast Time Delay Neural Networks Using Cross Correlation Performed in the Frequency Domain,” Proc. of IEEE World Congress on Computational Intelligence, IJCNN’06, Vancouver, BC, Canada, July 16-21, 2006, pp. 4990-4997.

[74] Hazem M. El-Bakry, and Nikos Mastorakis, “A Novel Model of Neural Networks for Fast Data Detection,” Proc. of the 7th WSEAS International Conference on Neural Networks, Cavtat, Croatia, June 12-14, 2006, pp. 144-151.

[75] Hazem M. El-Bakry, and Nikos Mastorakis, “A New Approach for Fast Face Detection,” Proc. of the 7th WSEAS International Conference on Neural Networks, Cavtat, Croatia, June 12-14, 2006, pp. 152-157.

[76] Hazem M. El-Bakry, "Pattern Detection Using Fast Normalized Neural Networks," Lecture Notes in Computer Science, Springer, vol. 3696, September 2005, pp. 447-454.

[77] Hazem M. El-Bakry, "Human Face Detection Using New High Speed Modular Neural Networks," Lecture Notes in Computer Science, Springer, vol. 3696, September 2005, pp. 543-550.

[78] Hazem M. El-Bakry, and Qiangfu Zhao, "Fast Pattern Detection Using Normalized Neural Networks and Cross Correlation in the Frequency Domain," EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Intelligent Vision Systems: Methods and Applications—Part I, vol. 2005, no. 13, 1 August 2005, pp. 2054-2060.

[79] Hazem M. El-Bakry, and Qiangfu Zhao, "Fast Time Delay Neural Networks," International Journal of Neural Systems, vol. 15, no.6, December 2005, pp. 445-455.

[80] Hazem M. El-Bakry, and Qiangfu Zhao, "Speeding-up Normalized Neural Networks For Face/Object Detection," Machine Graphics & Vision Journal (MG&V), vol. 14, No.1, 2005, pp. 29-59.

[81] Hazem M. El-Bakry, and Qiangfu Zhao, "A New Technique for Fast Pattern Recognition Using Normalized Neural Networks," WSEAS Transactions on Information Science and Applications, issue 11, vol. 2, November 2005, pp. 1816-1835.

[82] Hazem M. El-Bakry, and Qiangfu Zhao, "Fast Complex Valued Time Delay Neural Networks," International Journal of Computational Intelligence, vol.2, no.1, pp. 16-26, 2005.

[83] Hazem M. El-Bakry, and Qiangfu Zhao, "Fast Pattern Detection Using Neural Networks Realized in Frequency Domain," Enformatika Transactions on Engineering, Computing, and Technology, February 25-27, 2005, pp. 89-92.

[84] Hazem M. El-Bakry, "A New High Speed Neural Model For Character Recognition Using Cross Correlation and Matrix Decomposition," International Journal of Signal Processing, vol.2, no.3, 2005, pp. 183-202.

[85] Hazem M. El-Bakry, and Qiangfu Zhao, "Face Detection Using Fast Neural Processors and Image Decomposition," International Journal of Computational Intelligence, vol.1, no.4, 2004, pp. 313-316.

[86] Hazem M. El-Bakry, and H. Stoyan, "FNNs for Code Detection in Sequential Data Using Neural Networks for Communication Applications," Proc. of the First International Conference on Cybernetics and Information Technologies, Systems and Applications: CITSA 2004, pp. 21-25.

[87] Hazem M. El-Bakry, "Face detection using fast neural networks and image decomposition," Neurocomputing Journal, vol. 48, 2002, pp. 1039-1046.

[88] Hazem M. El-Bakry, "Human Iris Detection Using Fast Cooperative Modular Neural Nets and Image Decomposition," Machine Graphics & Vision Journal (MG&V), vol. 11, no. 4, 2002, pp. 498-512.

[89] Hazem M. El-Bakry "Fast Iris Detection for Personal Verification Using Modular Neural Networks," Lecture Notes in Computer Science, Springer, vol. 2206, October 2001, pp. 269-283.

[90] Hazem M. El-Bakry, "Automatic Human Face Recognition Using Modular Neural Networks," Machine Graphics & Vision Journal (MG&V), vol. 10, no. 1, 2001, pp. 47-73.

[91] http://www.storageswitzerland.com/Arcles/Entries/2011/6/16_Designing_Big_Data_Storage_Infrastructures.html

[92] “Switched Fabric” (Wikipedia), http://en.wikipedia.org/wiki/Switched_fabric, accessed May 27, 2014.

[93] “Big Data” (Wikipedia), http://en.wikipedia.org/wiki/Big_data

[94] “big data” http://www.webopedia.com/TERM/B/big_data.html

[95] “Big Data“ Industry Report 2014 IMEXResearch.com

[96] “NextGen Infrastructure for Big Data 2012 Storage Networking Industry Associa on”, Joseph White, Anil Vasuedeva “10 emerging technologies for Big Data” interviewed Dr. Satwant Kaur about the 10 emerging technologies that will drive Big Data forward December 4, 2012 .

[97] “Leveraging Hadoop-Based Big Data Architectures for a Scalable, High-Performance Analytics Platform” 2000488-001-EN Sept 2012 2012 Juniper Networks.

[98] “Making the Most of Big Data” Dr. Hossein Eslambolchi, former CTO of AT&T, is chairman and CEO of 2020 Venture Partners.

[99] “Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure”, by Y.Demchenko, P.Membrey, C.Ngo, C. de Laat, D.Gordijenko Submitted to Secure Data Management (SDM’13) Workshop. Part of VLDB2013 conference, 26-30 August 213, Trento, Italy

[100] “Big Data: What It Means for Data Center Infrastructure”, Krishna Kallakuri owner and vice president of DataFactZ SEPTEMBER 5, 2013

[101] “Big Data, Big Service“ 7th July 2013 by Tony Shan

"Big Data Solution Offering". MIKE2.0. Retrieved 14 May 2014.

[102] "Big Data Definition". MIKE2.0. Retrieved 9 March 2013.

[103] Boja, C; Pocovnicu, A; Bătăgan, L. (2012). "Distributed Parallel Architecture for Big Data".Informatica Economica 16 (2): 116–127.

[104] Klette R., and Zamperon, "Handbook of image processing operators, " John Wiley & Sonsltd, 1996.

[105] Cooley, J. W. and Tukey, J. W., "An algorithm for the machine calculation of complex Fourier series," Math. Comput. 19, 1965, pp. 297–301

Table 1. The theoretical speed up ratio (n=400).

Length of input data	Number of computation steps required for TNNs	Number of computation steps required for FNNs	Speed up ratio
10000	2.3014e+008	4.2926e+007	5.3613
40000	0.9493e+009	1.9614e+008	4.8397
90000	2.1478e+009	4.7344e+008	4.5365
160000	3.8257e+009	8.8219e+008	4.3366
250000	5.9830e+009	1.4275e+009	4.1912
360000	8.6195e+009	2.1134e+009	4.0786
490000	1.1735e+010	2.9430e+009	3.9876
640000	1.5331e+010	3.9192e+009	3.9119

Table 2. The theoretical speed up ratio (n =625).

Length of input data	Number of computation steps required for TNNs	Number of computation steps required for FNNs	Speed up ratio
10000	3.5132e+008	4.2919e+007	8.1857
40000	1.4754e+009	1.9613e+008	7.5226
90000	3.3489e+009	4.7343e+008	7.0737
160000	0.5972e+010	8.8218e+008	6.7694
250000	0.9344e+010	1.4275e+009	6.5458
360000	1.3466e+010	2.1134e+009	6.3717
490000	1.8337e+010	2.9430e+009	6.2306
640000	2.3958e+010	3.9192e+009	6.1129

Table 3. The theoretical speed up ratio (n =900).

Length of input data	Number of computation steps required for TRNNs	Number of computation steps required for FNNs	Speed up ratio
10000	4.9115e+008	4.2911e+007	11.4467
40000	2.1103e+009	1.9612e+008	10.7600
90000	4.8088e+009	4.7343e+008	10.1575
160000	0.8587e+010	8.8217e+008	9.7336
250000	1.3444e+010	1.4275e+009	9.4178
360000	1.9381e+010	2.1134e+009	9.1705
490000	2.6397e+010	2.9430e+009	8.9693
640000	3.4493e+010	3.9192e+009	8.8009

Table 4. Practical speed up ratio.

Length of input data	Speed up ratio (n=400)	Speed up ratio (n=625)	Speed up ratio (n=900)
10000	8.94	12.97	17.61
40000	8.60	12.56	17.22
90000	8.33	12.28	16.80

160000	8.07	12.07	16.53
250000	7.95	17.92	16.30
360000	7.79	11.62	16.14
490000	7.64	11.44	16.00
640000	7.04	11.27	15.89

Application of Artificial Intelligence on Classification of Attacks in IP Telephony

J. Safarik, M. Voznak, F. Rezac, J. Slachta

***Abstract*—**The paper deals with classification of attacks in IP telephony based on the multilayer perceptron neural network. Current analysis of these attacsk is typically based on statistical methods such as Hellinger-Distance, Holt-Winters or Brutlag's algorithm. The proposed solution MLP NN in the paper is used as a classifier of attacks in a distributed monitoring network of independent honeypot probes. 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 network and also information about implementation of this network. The trained neural network is capable to classify the most common used VoIP attacks. With the proposed approach is possible to detect malicious behavior in a different part of networks, which are logically or geographically divided and use the information from one network to harden security in other networks.

***Keywords*—** attack classification, multilayer perceptron network, neural network, SIP attacks.

I. Introduction

THE SIP (Session Initiation Protocol) is an open-source protocol which enables to establish, modify or terminate a general session [1], [2]. The IP telephony infrastructure based on SIP is very fragile to different types of attacks because of its similarity with HTTP and SMTP protocols. This can lead to loss of money, trust and other unpleasant consequences [3].

This situation could be solved partially with strict security rules, encryption and properly set VoIP servers. Even then is an attacker able to corrupt whole IP telephony network and stole sensitive information, eavesdrop calls, stole caller identities or deny the service (DoS attack).

Monitoring of VoIP infrastructure, IDS/IPS (Intrusion Detection/Prevention Systems) or honeypots application can detect these attacks and malicious activity in the network. 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.

With an automatic classification system is possible to automatically detect attacks on IP telephony from various set of honeypots and harden existing security mechanism. This honeypot network concept is closely described in the following section.

II. State of the Art

Detection of SIP infrastructure attack is solved with different ways in a range of studies and papers. Some methods rely on IDS system as Snort and its features or implement new features for better attack detection [4]. Other ways use statistical methods to analyze attributes of the SIP traffic [5]. There are methods for attack recognition based on Hellinger-Distance [6], forecast methods like Holt-Winters [7] and Brutlag's algorithm [8] or variety of SIP traffic anomaly detections.

Hellinger-Distance is used to quantify the similarity between two probability distributions (1), where p is the distribution of data within training period and q the distribution of data within short period [6].

$$H^2(P,Q)=\frac{1}{2}\sum_{i=1}^{n}\left(\sqrt{p_i}-\sqrt{q_i}\right)^2 \qquad (1)$$

Holt-Winters model is used for detection of anomaly using predictive approach, see relations (2) – (5), this model is called also as the triple exponential smoothing model and it is a well-known adaptive model used to modeling time series characterized by trend and seasonality [7].

$$\hat{y}_t = L_{t-1} + P_{t-1} + S_{t-T} \qquad (2)$$

where L is a level component given by :

This research has been supported by the Ministry of Education of the Czech Republic within the project LM2010005..

J. Safarik is a PhD. student with Dept. of Telecommunications, Technical University of Ostrava and he is also a researcher with Dept. of Multimedia in CESNET, Zikova 4, 160 00 Prague 6, Czech Republic (e-mail: safarik@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.

F. Rezac is a PhD. student with Dept. of Telecommunications, Technical University of Ostrava and he is also a researcher with Dept. of Multimedia in CESNET, Zikova 4, 160 00 Prague 6, Czech Republic (e-mail: rezac@cesnet.cz).

J. Slachta is a M.S. student with Dept. of Telecommunications, Technical University of Ostrava and he is also a researcher with Dept. of Multimedia in CESNET, Zikova 4, 160 00 Prague 6, Czech Republic (e-mail: slachta@cesnet.cz).

$$\mathrm{L_t} = \alpha(y_t - S_{t-T}) + (1-\alpha)(L_{t-1} + P_{t-1}) \qquad (3)$$

P is trend component given by :

$$Pt = \beta(L_t - L_{t-1}) + (1-\beta)P_{t-1} \qquad (4)$$

And S is seasonal component given by :

$$St = \gamma(y_t - L_t) + (1-\gamma)S_{t-T} \qquad (5)$$

Holt-Winters method was used to detect network traffic anomalies by Brutlag in [8]. In his concept of confidence bands, parameters $\hat{y}_{\max_t}$ and $\hat{y}_{\min_t}$ were introduced and is possible to measure deviation for each time point in the seasonal cycle (6).

$$\begin{aligned} \hat{y}_{\max_t} &= L_{t-1} + P_{t-1} + S_t - T + m * d_{t-T} \\ \hat{y}_{\min_t} &= L_{t-1} + P_{t-1} + S_t - T - m * d_{t-T} \end{aligned} \qquad (6)$$

Where d is predicted deviation given by (7):

$$d_t = \gamma \,|\, y_t - \hat{y}_t \,| + (1-\gamma)d_{t-T} \qquad (7)$$

Where $\hat{y}_t$ is predicted value of variable in moment t and y_t is measured value of variable in moment t and T is time series period. Then α is data smoothing factor, β is trend smoothing factor and γ is the seasonal smoothing factor, finally m is the scaling factor for Brutlags confidence bands.

This paper proposes a SIP attack classification with MLP (multilayer perceptron) neural network from honeypot application Dionaea, the alogrithm is described in chapter IV.

III. Honeypot Network Concept

The classification engine based on MLP network is only a part of IP telephony infrastructure protection. A single honeypot application brings valuable information, with a combination of honeypots running in different networks at different locations should provide even more detailed data.

Exceeding number of running honeypots causes higher requirements for their management and support. The proposed design of a distributed honeypot network, shown at Fig. 1, solves this problem with a centralized server for data gathering, analysis and honeypot monitoring.

The main part of distributed network concept is honeypot image, which contains already prepared and preconfigured honeypot application, operation system and other software needed. This single node communicates with a centralized server via secured channel and periodically sends information about detected attacks to server.

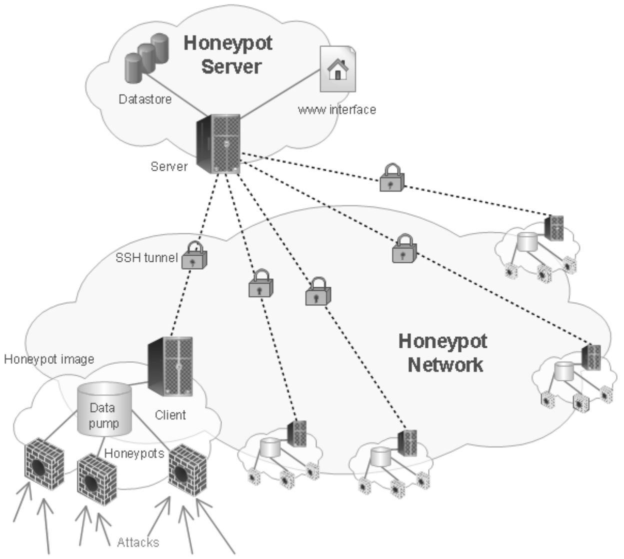


Fig. 1 Honeypot network concept

Neural network algorithm described in this paper is used in honeypot network hierarchy as a module on the centralized server for classification of SIP based attacks. More information about distributed honeypot network covers previous paper [9].

IV. MLP Neural Network

Neural networks try to model information processing capabilities of the nervous system of mammals. This nervous system is composed of millions of interconnected cells in a complex arrangement. The artificial neural network tries to model this design. The function of a single neuron is well known and serves as a model for an artificial one.

Even that we do not complete understand the complexity and massive hierarchical networking of the brain, with its incredible processing rate, artificial neural network handle complex problems by using different topologies. Many versions of these topologies are known today, each of them has its pros and cons.

The feed-forward MLP neural network was used for VoIP attack classifications. It consists of several layers, each containing the specific number of neurons called perceptron. These perceptrons in one layer are interconnected to each other in the following layer (this connection could be also called a synapse) [10].

The Fig. 2 shows the inner structure of used MLP network. The MLP network solution used for classification has two hidden layers, with one input and one output layer. Each neuron in the input layer has a value based on input parameters.

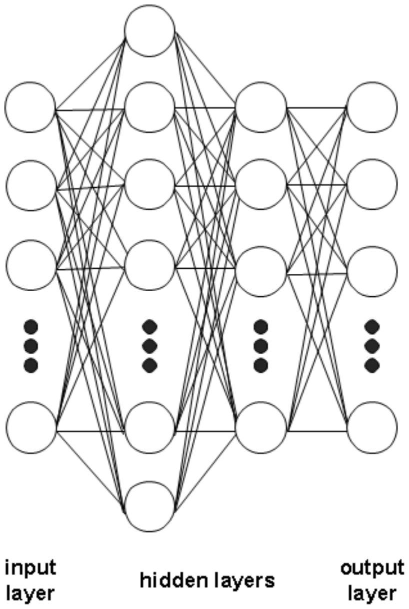


Fig. 2 MLP neural network topology

This layer has the same number of neurons as there are parameters in the input set. After the input layer continues two hidden layers and output layer. The output layer has the same number of neurons as the number of attack classes, so each neuron is then a single class of learned attack. Number of neurons inside hidden layers depends on neural network configuration and are typically higher than the number of neurons in input or output layers.

A. Perceptron

The perceptron is a more general computational model than McCulloch-Pitts units. The main innovation is in including numerical weights for connections and a special interconnection pattern. The activation function for neuron – sigmoid impacts the final potential of a neuron. This result potential is then transmitted through connections to neurons in the next layer while afflicted by each connection weight. These weights serve as a memory for neural network. Inputs for the activation function are real inputs x_1, x_2,…, x_n from the previous layer, with the associated connection weights w_1, w_2,…, w_n.

The output of a neuron is between *0* and *1*, where *0* means inhibition and *1* excitation. The final value on the output of neuron (y) depends on its activation function. As was mentioned before, this function is a real sigmoid function (8) and (9).

$$S_C : \Re \rightarrow (0,1) \tag{8}$$

$$y = S_C(z) = \frac{1}{1+e^{-cz}} \tag{9}$$

The relation (10) shows parameter z, which is the sum of the output from previous layer neuron x and multiplies by corresponding connection weight w. Parameter c represents a skewness of the sigmoid function (typically it is *1.0*). Higher values of parameter c bring the skewness of a sigmoid function closer to the step function [11], [12].

$$z = \sum_{i=1}^{n} w_i x_i \tag{10}$$

B. Backpropagation Algorithm

As was mentioned before, the memory of neural network is saved in connection weights. The neural network learning mechanism – backpropagation is used to acquire these values. While classifying, the neural network is feed-forward mode and information is transferred from the input layer to the output layer. Backpropagation works as a reverse mechanism to feed-forward, with the specific set of data called training set. Training set has same the format as attack inputs for neural networks but contains also the final result of classification (or the class of the specific attack).

The core of a backpropagation algorithm and the neural network learning is a process of weight adaptation. It is done on the training set of inputs with known outputs. The solution of learning problem is a combination of weights with the minimal error function. Learn rate parameter (η) affects connection weight correction, used to lower the value of the error function (13).

$$\delta_j = \sum_{k=1}^{n} \delta_k y_k (1 - y_k) c w_{jk} \tag{11}$$

The equation (11) shows computation the of backpropagation error (δ) for connection weight in one layer (indexed as j). It is counted as a multiplication of higher layer (indexed as k) backpropagation error, actual output, expected output and actual weight of the connection. Parameter y represents the output of neuron, x always its inputs. Parameter c is the expected output and w the connection weight. The backpropagation error is then used in weight adaptation equations (12) and (13).

$$\Delta w_{ij} = \eta \delta_j y_i \tag{12}$$

$$w_{ij} = w_{ij} + \Delta w_{ij} \tag{13}$$

The learn rate parameter (η) serves to set a proper step of correction in one backpropagation iteration [11], [12]. One iteration of backpropagation learning uses all records from the training set. The last parameter w_{ij} is the connection weight from the previous layer (i) to the actual layer (j) as shown Fig. 3.

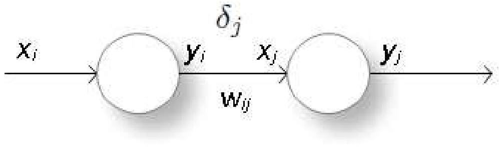


Fig. 3 Indexing between layers

V. Practical Implementation

In previous research covered in [12] was used MLP neural network for detection of six basic SIP attacks. Because there are changes in input vector and attack classes, inner structure of MLP network was changed.

A. MLP neural network configuration

The new final neural network contains 10 input layer neurons as the previous generation. The two hidden layers contain 30 and 24 neurons, the last and output layer 8 neurons. The previous generation was not able to detect correctly one class of attacks, so it detects only five types of SIP attack. With the change to eight classes, there is more robust and accurate detection of attacks.

The inner structure of previous generation network was based on test of convergence of 100 backpropagation iterations. These tests proved different mean times of learning for different inner structures (see Fig. 4 (timesXXYY – 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.

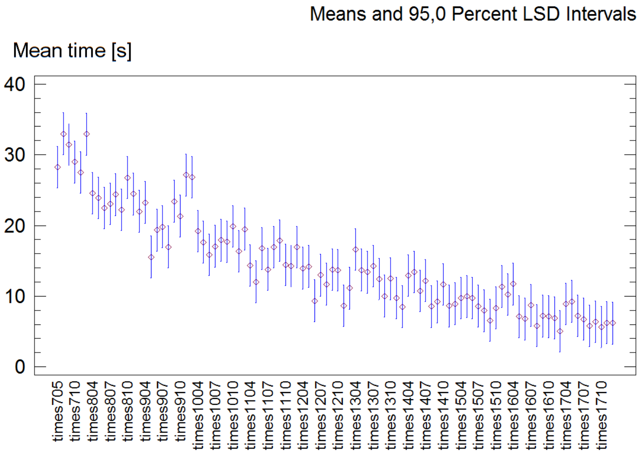


Fig. 4. Mean times of 100 backpropagation iteration with different inner structure

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.

Both generations of MLP neural networks are learned to the same confidence interval. The successfulness on the training set is always lower than 5%. This ensures statistically significant classification capability on the training set. Both generations use the same configuration of skewness (*1.0*) and learn momentum (*0.8*).

B. Data source for classification & input vector parameters

All attack information is gathered through multi-service oriented honeypot application Dionaea. We choose Dionaea for its features tested in previous research. It emulates and monitors traffic of a SIP PBX (Private Branch eXchange). However only specific set of information is saved to the internal database e.g. used SIP message, IP addresses, ports or specific SIP header values.

Dionea honeypot contains strict information about malicious traffic. All running honeypots are accessible through internet on public IP addresses (IPv4). No legitimate calls or devices connect to this honeypot, so only malicious traffic or misconfigured devices communicate with it.

All attack data save Dionaea to sqlite database. The database contains several tables for specific protocols and functions. But all tables have a single pointer to a table connection, which contains basic information about attack. Information about SIP attacks is distributed in five tables, each with different information. Selecting only single lines from these tables for classification is valueless, partly because of request/response behavior of SIP protocol. All data for final classification are aggregated from selected tables to an array with 10 attributes.

These individual 10 attributes then serve as an attack vector (or neural network input). The aggregation depends on attack origin and also time of last message occurrence (there is 5 minute sliding window after last message detection). Attributes are in the following order: attack time duration; connection count; REGISTER message count; INVITE msg. count; ACK msg. count; BYE msg. count; CANCEL msg. count; OPTIONS msg. count; SUBSCRIBE msg. count; connection rate. The connection count attribute holds the number of connection from a single source on honeypot. The connection rate is the ratio of all received SIP messages to connection count.

C. Backpropagation configuration

MLP neural network use backpropagation algorithm for learning. SIP attack classification 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%). Before backpropagation starts, all connection weights are set randomly from range (*-1,1*). After first initialization of weights starts backpropagation iteration. Each iteration cycle uses all items from the training set. After specific number of iteration cycles (100) is automatically checked successfulness of classification. If the successfulness on the training set is lower than specified threshold, backpropagation runs another iteration cycle. When the successfulness is higher than 0.95, the neural network learning is done. To avoid the possibility of stuck in local extreme, system automatically reinitialize all connection weights after 2 500 000 backpropagation cycles.

D. Training set

The training set, along with neural network input, is one of key parts of neural network. If the items in the training list are not specific representative of an attack group, the attack cannot be successfully classified. This misclassification has a negative influence on learning and prevents successful learning of attacks.

As a source for the training set was always used real attack traffic. This traffic is aggregated and then classified by hand. From this classified set is prepared training set with attack groups. The new MLP neural network training set consist of 104 items, 13 items for each attack group. These classes are options tests; options scanning; call testing; unknown protocol; register and call; registration test, registration flooding; register attempt. This set of attack classes corresponds to detected types of attack from a period of two months.

E. Reference set

As a reference set serves an aggregated set attack vector detected on various honeypots. All runs Dionaea application but on different hardware, IP addresses and in different geographical locations. One honeypot is masquerading behind a set of IP addresses for raising the malicious traffic. Each honeypot application runs for other period because they are not started together, so they provide a diverse group of detected attacks. All attacks detected on these honeypots were classified by hand, so we cannot eliminate human factor error.

Tests of new MLP neural network on three reference data set bring exciting information. These set do not contain data from the training set. Result of analyses with MLP networks has following successfulness: 94.94%; 79.85% and 97.54%. Totally were detected 1631 attack groups and 57752 SIP messages (data for three months period). The lowest classification precision 79.85% was caused by new call attack, which was not included in the training set.

VI. Conclusion

SIP protocol is an open-source protocol and becomes one of the most used protocols for handling of VoIP services. There is even estimated rise of SIP devices in the future and the security of SIP device and PBX will become a crucial question. This situation will lead to a higher exposition to various types of attacks and even misconfigured devices. These factors will have a negative impact on VoIP service. Previous research in our laboratory confirmed high vulnerability of SIP servers to various attacks [13].

The proposal distributed honeypot network in combination with neural network classifiers serves as another security level. But the potential lies not only in detection capability and attack research. With the possibility to change firewall rules or network routing, whole system can prepare precaution mechanisms against attacks even when it do not influence the target network. The proposal of such monitoring system is distributed honeypot network.

As we found out, similar research on using neural networks with for identification and classification of attacks in VoIP has been carried out in [14], where a feedforward artificial neural network was applied as well. Nevertheless, this research was focused mainly on identification of DoS attacks and can not be compared with our results. Since we exploit data from honeypots, where only malicious traffic is directed, our research is strictly oriented on classification of attacks. The trained classifier was able to distinguish the particular type of attack with high reliability.

Classification by human is very precise, but time consuming and expensive. It is typically conducted after the damage is done. Automatic classification mechanism brings a solution for VoIP classification and simplifies the analysis of attacks. The biggest disadvantage of this solution is it strong bindings on the training set. The MLP neural network cannot adapt to new attack classes or scenarios.

Test of the new MLP classification network on reference sets prove its quality but shows its limits and new ways for improvements. The future plans for neural network classification cover improving the accuracy of existing solutions and implementation of other evolutionary and statistical algorithms for attack classification. One of challenges is also in detection of attacks in legitimate VoIP traffic.

References

[1] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, “SIP: Session Initiation Protocol”, IETF RFC 3261, June 2002.

[2] L. Macura, M. Voznak, K. Tomala, J. Slachta, "Embedded multiplatform SIP server solution", In Proc. *TSP 2012*, art. no. 6256295, 2012, pp. 263-266.

[3] F. Rezac, M. Voznak, J. Ruzicka, “Security Risks in IP Telephony”, In Proc. *Cesnet Conference 2008-Security*, 2008, pp. 31-38.

[4] J. Gomez, C. Gil, N. Padilla, R. Banos, C. Jimenez, ”Design of a Snort-Based Hybrid Intrusion Detection System”, In *Lecture Notes in Computer Science*, *Volume 5518, 2009*, pp 515-522.

[5] H. J. Kang, Z. Zhang, S. Ranjan, A. Nucci, “Sip-based VoIP Traffic Behavior Profiling and Its Applications”, In Proc. *MineNet’07*, 2007.

[6] H. Sengar, H. Wang, D. Wijesekera, S. Jajodia, “Detecting VoIP Floods Using the Hellinger Distance”, In *IEEE transactions on parallel and distributed systems, Vol. 19, No. 6*, June 2008.

[7] M. Szmit, A. Szmit, "Usage of Modified Holt-Winters Method in the Anomaly Detection of Network Traffic: Case Studies, In *Journal of Computer Networks and Communications, Volume 2012, 2012*, Article ID 192913.

[8] J. D. Brutlag, "Aberrant Behavior Detection in Time Series for Network Monitoring", In Proc. *14th System Administration Conference*, New Orleans, 2000, pp. 139-146.

[9] J. Safarik, J., Voznak, M., Rezac, F., Partila, P., Tomala, K., “Automatic Analysis of Attack Data from Distributed Honeypot Network”, In Proc. *Mobile Multimedia/Image Processing, Security, and Applications 2013*, 875512, Baltimore, May 28, 2013.

[10] R. Rojas, *Neural Networks*, Springer-Verlag, 1996, ISBN 978-3540605058.

[11] J. Heaton, *Introduction to Neural Networks for JAVA*, 2nd Edition”, Heaton Research, 2008, ISBN 978-1604390087.

[12] J. Safarik, P. Partila, F. Rezac, L. Macura, M. Voznak, “Automatic Classification of Attacks on IP Telephony”, In *Advances in Electrical and Electronic Engineering, Vol. 11, Issue 6*, 2013, pp. 481-486, ISSN 1336-1376.

[13] F. Rezac, M. Voznak, K. Tomala, J. Rozhon, J. Vychodil, “Security Analysis System to Detect Threats on a SIP VoIP Infrastructure Elements”, In *Advances in Electrical and Electronic Engineering, Vol. 9, No. 5, 225-232*, 2011, ISSN 1336-1376.

[14] N. Shekokar and S. Devane, "Anomaly detection in VoIP system using neural network and fuzzy logic," In *Communications in Computer and Information Science*, Volume 250 CCIS, 2011, pp. 537-542.

A New 2D Image Compression Technique for 3D Surface Reconstruction

M. M. Siddeq, Prof. M. Rodrigues

Abstract— Image compression is one of the important techniques used today for image and video transmission. There are many types of image compression techniques are used these days; one of them is JPEG technique. In this research, we introduce a new idea for applying the JPEG technique with Discrete Wavelet Transform (DWT) for high-resolution images. Our image compression algorithm consists of; firstly, transform an image by single level DWT. Secondly, JPEG algorithm applied on "LL" sub-band this process is called JPEG Transformation. Thirdly, separate the final transformed matrix into DC-Array and AC-Matrix contains DC values and AC coefficients respectively. Finally, the minimize-matrix-size algorithm applied on AC-Matrix followed by arithmetic coding. The novel decompression algorithm used in this research is Parallel Sequential Search Algorithm, which is represented inverse minimize-matrix-size algorithm. The searching algorithm consist of a P pointers, all these pointers are working in parallel to find the original AC-coefficients. Thereafter, combines all decoded DC-values with the decoded AC-coefficients in one matrix followed by apply inverse JPEG transformed and inverse DWT. the technique is tested by compression and reconstruction of 3D surface patches. Additionally, this technique is compared with JPEG and JPEG2000 algorithm by using 2D and 3D RMSE.

Keywords—DWT, JPEG, Minimize-Matrix-Size Algorithm, Parallel SS-Algorithm, 3D reconstruction

I. Introduction

Compression methods are being rapidly developed to compress large data files such as images, where data compression in multimedia applications has lately become more vital. With the increasing growth of technology and the entrance into the digital age, a vast amount of image data must be stored in a proper way using efficient methods usually succeed in compressing images, while retaining high image quality and marginal reduction in image size. Since first attempts, the discrete cosine transform (DCT) domain has been used **[1]**. Image is divided into segments and each segment is then a subject of the transform, creating a series of frequency components that correspond with detail levels of the image. Several forms of coding are applied in order to store only coefficients that are found as significant. Such a way is used in the popular JPEG file format, and most video compression methods and multi-media applications are generally based on it **[2][4]**.

A step beyond JPEG is the JPEG2000 that is based on wavelet transform which is one of the mathematical tools for hierarchically decomposing functions. Image compression using Wavelet Transforms is a powerful method that is preferred by scientists to get the compressed images at higher compression ratios with higher PSNR values **[3][5]**. Its superiority in achieving high compression ratio, error resilience, and other features promotes it to become the tomorrow's compression standard and leads to the JPEG2000 ISO. As referred to the JPEG abbreviation which stands for Joint Photographic Expert Group, JPEG2000 codec is more efficient than its predecessor JPEG and overcomes its drawbacks **[12]**. It also offers higher flexibility compared to even many other codec such as region of interest, high dynamic range of intensity values, multi component, lossy and lossless compression, efficient computation, compression rate control, etc. The robustness of JPEG2000 stems from its utilization to the Discrete Wavelet Transform (DWT) in encoding the image data. DWT exhibits high effectiveness in image compression due to its support to multi-resolution representation in both spatial and frequency domains. In addition, DWT supports progressive image transmission and region of interest coding **[13][14].**

II. Proposed 2DImage Compression Algorithm

JPEG technique is one of the greatest techniques are used in the image compression; also it is used for encryption and steganography. The important feature of the JPEG it is uses the "Quality" parameter, which allow for the user to adjust the amount of the data lost over a very wide range. In this section we explain in details about JPEG transformation applied on the discrete wavelet transform. The JPEG transformation consists of; 1) Apply DCT on each 8x8 block followed by quantization process. 2) Zigzag scan used for converting each block into 64 coefficients, and store 64-coefficients in two different matrices **[11]**. Fig.-1 describes the proposed DWT-JPEG algorithm steps.

M. M. Siddeq, M. A. Rodrigues, Geometric Modeling and Pattern Recognition Research Group, Sheffield Hallam University, Sheffield, UK
e-mail: mamadmmx76@yahoo.com , M.Rodrigues@shu.ac.uk

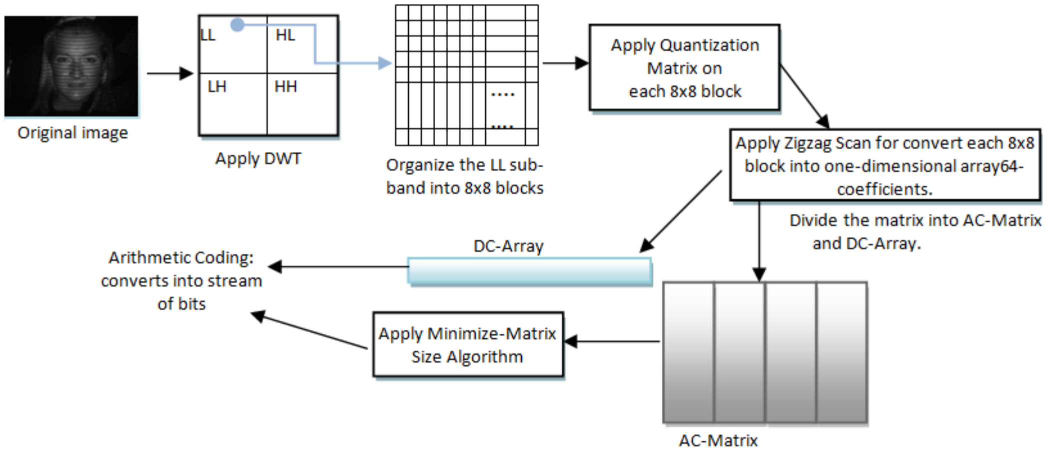

Original image
LL
HL
LH
HH
Apply DWT
....
....
Organize the LL sub-band into 8x8 blocks
Apply Quantization Matrix on each 8x8 block
Apply Zigzag Scan for convert each 8x8 block into one-dimensional array64-coefficients.
Divide the matrix into AC-Matrix and DC-Array.
DC-Array
Arithmetic Coding: converts into stream of bits
Apply Minimize-Matrix Size Algorithm
AC-Matrix

Fig.-1, Proposed DWT-JPEG Compression Technique

A. Discrete Wavelet Transform

DWT is the first phase in the proposed image compression algorithm, to produces four sub-bands (See Figure-1). The top-left corner is called "LL", represents low-frequency coefficients, and the top-right called "HL" consists of residual vertical frequencies. The bottom-left corner "LH", and bottom-right corner "HH" are represents; residual horizontal and residual vertical frequencies respectively [5]. Most values in the high-frequency domains (i.e HL, LH and HH) are zeros or insignificant coefficients without affecting on the reconstructed image. For this reason all the high frequency domains are discarded in this research (i.e. set all values to zero), and this does not mean the image will lose much information, this is depends on the image dimensions. DWT uses filters for decomposing image [15], these filters assists to increase the number of zeros in high frequency sub-bands. One of these filters is used in decomposition and composition is called Daubechies Filter. This filter stores much information about the original image in the "LL" sub-band, while other high-frequency domains contain less significant details, and this is one of important property in Daubechies filter. The reconstructed image just need "LL" sub-band, while other high-frequency sub-bands are omitted, and this is the key for achieving a high compression ratio [8][9]. Fig.-2 shows the decomposition image by Daubechies filter, and then recomposes sub-bands without high-frequencies.

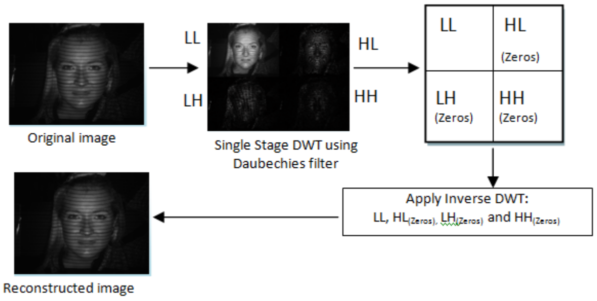


Fig.-2 reconstructed image by using Daubechies single stage DWT

B. JPEG Transformation

The "LL" sub-band partitioned into non-overlapped 8x8 blocks, each block is transformed by using two-dimensional DCT to produce de-correlated coefficients. Each 8x8 frequency domain consists of; DC-value at the first location, while other coefficients are called AC coefficients [5].

After applying the two-dimensional DCT on each 8x8 block, each block quantized by the Quantization Matrix "QM" using dot-division-matrix with truncates the results. This process removes insignificant coefficients and increases the number of the zeros in the each block. QM computes as follows:

$$QM(i,\ j) = \begin{cases} Block + (i + j), & if\,(QM(i,\ j)) = odd \\ Block + (i + j) + 1 & if\ (QM(i,\ j)) = even \end{cases} \quad (1)$$

Where \ Block: is represented block size
i ,j=1,2,3…,Block

$$QM(i, j) = QM(i, j) * Scale \qquad (2)$$

In the above Eq. (2), the factor "Scale" it is used to increase/decrease the values of the "QM". Thus, image details is reduced in case of the factor Scale >1. There is no limited range for this factor, because this is depends on the DCT coefficients.

Each quantized 8x8 block is converted into one-dimensional array (i.e. the array contains 64 coefficients) by zigzag scan [13].Whereas, the first value transferred into new array called DC-Array, while others are 63 coefficients are stored to new matrix "LL_{AC}". Finally, the DC-Array is compressed by using Arithmetic coding. The Arithmetic coding is one of the important methods used in data compression method, especially this method used in JPEG2000. Arithmetic coding depends on "Low" and "High" equations to generate streams of bits [5].

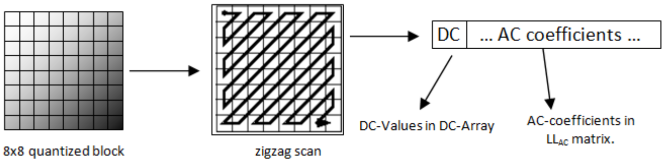
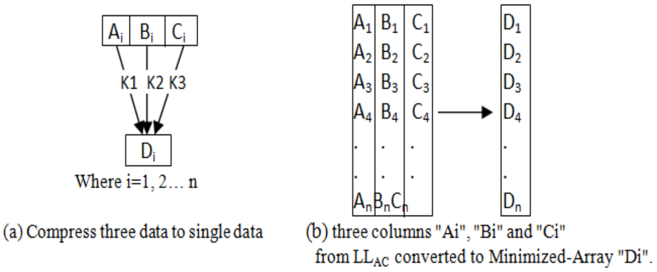


Fig.-3, operation and separate DC-value from64 coefficients

III. Minimize-Matrix-Size Algorithm

LL$_{AC}$ Matrix ready for coding by Minimize-Matrix-Size Algorithm, this algorithm applied on each three coefficients, to produce single data. This means reduce each three columns to single coded array which is called Minimized-Array. However, the bit size for each data in the Minimized-Array increased. Fig.-4 shows converting three columns into one dimensional array [9][10].



(a) Compress three data to single data

(b) three columns "Ai", "Bi" and "Ci" from LL$_{AC}$ converted to Minimized-Array "Di".

Fig.-4, Minimize-Matrix-Size Algorithm

In above figure-4 (a) K1, K2 and K3 represents key for conversion. The following equation illustrates converting three data, to single data.

$$D_i=(K_1 \times A_i)+(K_2 \times B_i) + (K_3 \times C_i) \qquad (3)$$

Where\ $i=1, 2, 3,\ldots n$

If the key is lost, the data cannot be retrieved, because the keys are used in coding and decoding. The key values generated through random number generator. For example, assume we have the following array: [3 -9 0], Maximum value in the array=|-9|=M= 9, and Base Value=0.1; Key1= 0.8147, Key2=0.9058, and Key3=0.1270. The key generated once for all matrix data, after calculation, all coded data "Di" arranged together to be one-dimensional array (i.e. Minimized-Array).

Before apply the Minimize-Matrix-Size algorithm, the algorithm computes the probability of the data for AC-matrix. These probabilities are called Limited-Data, which is used later in decompression stage [10]. The Limited-Data stored as additional information with compressed data. Fig.-5, describes Limited-Data computed from original matrix.

The Minimized-array contains positive and negative data, and each data size reached to 32-bit, these data can be compress by a coding method, but the index size (i.e. header compressed file) reaches to 50% of compressed data size. The index data are used in decompression stage, therefore, the index data breakup into parts for easy compress. Each data partitioned into parts: 4-bit (i.e. each data in index may be breakup into six 4-bit data), and this process increase the probability of redundant data, finally, coded by arithmetic coding.

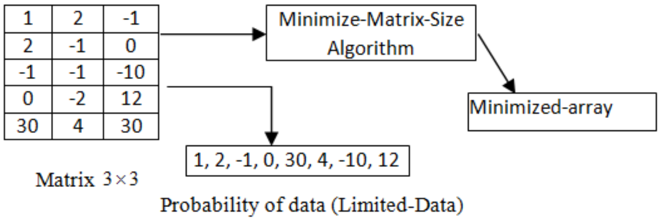


Fig.-5, Limited-Data

IV. Proposed Decompression Algorithm

The decompression algorithm represents reverse steps for the proposed image compression. Firstly, applied arithmetic decoding for decompress DC-array, Nonzero-array and Zero-array. Thereafter, nonzero-array and zero-array are combined together for reconstructing minimized-array. Secondly, using novel Parallel Sequential Search Algorithm (PSS-Algorithm), moreover, this algorithm represents inverse Minimize-Matrix-Size Algorithm for reconstructing AC-Matrix. PSS-Algorithm, estimates (Ai, Bi and Ci) by using "Di" with Key. Whereas, Ai, Bi and Ci are represents estimated columns for decompress AC-Matrix. PSS-Algorithm can be illustrates in the following steps:

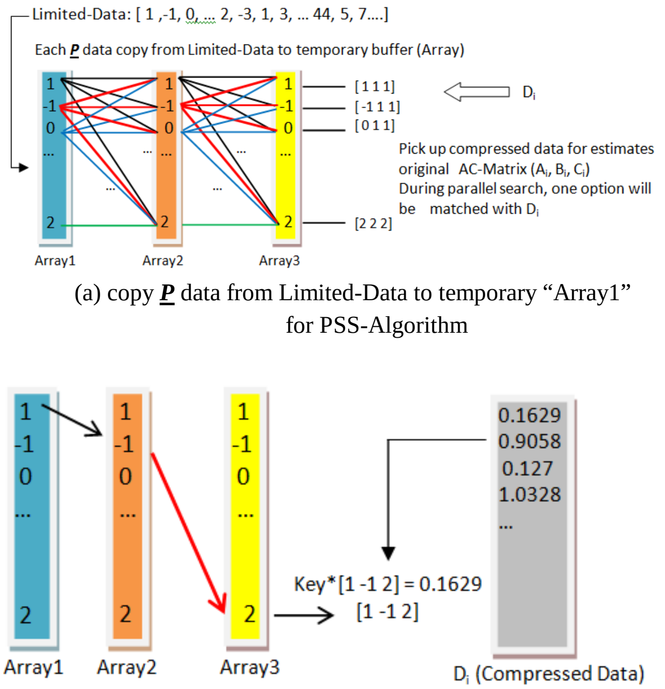
Step 1: PSS-Algorithm starts to pick first P data from the Limited-Data, and then these P data are connected with each other look like a network as shown in Fig.-6.

In Fig.-6 "Column-1" data connected as a network with "Column-2" data, also "Column-2" is networking with "Column-3". In another words, the searching algorithm computes all options in parallel. For example: A=[1 -1 0] , B=[1 -2 0] and C=[3 -1 5], and P=3, according to Eq.(3) "A","B" and "C" computes 27 times. This means, all options computes in parallel and one option will be matched with the "Di", and "Ai", "Bi" and "Ci" in "Column-1", "Column-2" and "Column-3" represented decompressed data.

Initially, PSS algorithm starts with P=10 data from "Limited-Data(1…10.)" that used by the algorithm, these data are estimates three columns (A, B and C), as mentioned in Figure-8(a). Thereafter, the algorithm starts searching for original data (Ai, Bi and Ci) which is depends on compressed column "Di" and Key-values. The first iteration for the algorithm starts with matching selected "Di" with 10 outputs from PSS-algorithm (i.e. P=10, three columns = P3= 1000 data). In another words ,Eq.(3) executed 1000 times in parallel for finding original values for columns (A,B and C) as mentioned in Figure-7(b). If result unmatched, in this case the second option will be taken form "Limited-Data(11…20.)" (i.e. selecting another 10 data from Limited-Data transferred to Array1), while "Array2" and "Array3" are remains in same old options, if the processing still did not find the result, in this case"Array2=Array1" (i.e. transferred data from Array1 to Array2), then new processing starts.

Through this explanation, "Array1", "Array2" and "Array3" are working like digital clock: sec, min. and hour respectively, this process will continue until finding all original columns (Ai, Bi and Ci) in AC-Matrix.

Step 2: In this step decompressed AC-Matrix composed with each DC-value (i.e. DC-values from DC-array), then followed by zigzag scan to convert each 64-coefficients to 8x8 blocks. These blocks combined with each other to build LL sub-bands. Subsequently, applied inverse quantization (i.e. dot-multiplication), followed by inverse DCT on each 8x8 block. Finally, applied inverse DWT for obtaining 2D image linked with 3D application for reconstruct 3D image. The decompression algorithm steps are showed in Figure-8.


Limited-Data: [1 ,-1, 0, ... 2, -3, 1, 3, ... 44, 5, 7....]
Each P data copy from Limited-Data to temporary buffer (Array)
1
-1
0
...
2
[1 1 1]
[-1 1 1]
[0 1 1]
[2 2 2]
Di
Pick up compressed data for estimates original AC-Matrix (Ai, Bi, Ci)
During parallel search, one option will be matched with Di
Array1
Array2
Array3
0.1629
0.9058
0.127
1.0328
...
Key*[1 -1 2] = 0.1629
[1 -1 2]
Di (Compressed Data)

(a) copy **P** data from Limited-Data to temporary "Array1" for PSS-Algorithm

(b) data matched through PSS-Algorithm

Fig.-7, (a, b) strategy work for PSS-Algorithm

V. Experiments Results

The proposed compression system applied on images, as shown in Figure-9. The tests have been performed using Daubechies DWT (db3), the block sizes used by DCT (4×4 and 8×8). The results described below used Matlab for 2D image compression in connection with a 3D visualization software running on an AMD Quad-Core microprocessor. Tables: 1 and 2 shows the compressed size possibilities for each image.

The proposed decompression algorithm applied on each compressed data, as mentioned before in section 3. The decompressed algorithm shows range of image quality according to "Scale" parameter and blocks size used in and JPEG-Transformation (See Eq.(2)). Figure-10 and Figure-11 shows sequence of decompressed 3D images: Face1 and FACE2 respectively.

Tables: 3 and 4 shows time execution for PSS-algorithm for each image using two types of pointers (**P**=5 and **P**=10). The pointer refers to number of coefficients using in parallel for space search (i.e. searching in Limited-Data).

Table-1, 2D image"FACE1.bmp" compressed by the proposed image compression algorithm

Scale – parameter used by quantization	Block size used by JPEG-Transformation	Compressed size
2	4x4	51.6 Kbytes
2	8x8	28.4 Kbytes
4	8x8	16.3 Kbytes
5	8x8	13.5 Kbytes
6	8x8	11.6 Kbytes

Table-2, 2D image"FACE2.bmp" compressed by the proposed image compression algorithm

Scale – parameter used by quantization	Block size used by JPEG-Transformation	Compressed size
2	4x4	33 Kbytes
2	8x8	16.8 Kbytes
4	8x8	9.4 Kbytes
5	8x8	7.7 Kbytes

Table 3, Parallel Search algorithm time execution for image: FACE1.bmp

Parameters	PSS-Algorithm, **P**=5		PSS-Algorithm, **P**=10
Scale –used by quantization	Block size	Total time(sec.)	Total time (sec.)
2	[4x4]	126.20	122.24
2	[8x8]	65.59	61.12
4	[8x8]	15.22	8.47
5	[8x8]	9.37	6.91
6	[8x8]	6.14	4.91
8	[8x8]	3.38	4.77

Table 4, Parallel Search algorithm time execution for image: FACE2.bmp

Parameters	PSS-Algorithm, **P**=5		PSS-Algorithm, **P**=10
Scale –used by quantization	Block size	Total time(sec.)	Total time (sec.)
2	[4x4]	16.27	10.67
2	[8x8]	9.0	7.78
4	[8x8]	3.1	3.77
5	[8x8]	3.0	3.21

VI. Comparison with JPEG, JPEG2000

Our approach is compared with JPEG and JPEG2000; these two techniques are used widely in digital image compression, especially for image transmission and video compression. The JPEG technique is based on the 2D DCT applied on the partitioned image into 8x8 blocks, and then each block encoded by RLE and Huffman encoding [4]. The JPEG2000 is based on the multi-level DWT 9/7-daubaches filter, applied on the partitioned image and then each partition quantized and coded by Arithmetic encoding. Most image compression applications allow the user to specify a quality parameter for the compression. If the image quality is increased the compression ratio is decreased and vice versa [5]. The comparison is based on the 2D image and 3D image for test the quality by Root-Mean-Square-Error (RMSE). Tables: 5 and 6 shows the comparison between three methods for Face1, Face2 respectively.

In tables: 5 and 6 "NON" refers JPEG algorithm unable to compress/decompress an image at high compression ratio, while other two methods (our proposed and JPEG2000) are able to compress/decompress successfully. In some cases the 3D RMSE vary, if we compare it with 2D RMSE, this is because the dimensions of original 3D image and 3D decompressed image unmatched. In this case the unmatched regions are discarded. On the other hand, RMSE is not enough to show the real comparison between these three methods. The following figures: 12 and 13 shows the visual properties for the 3D decompressed images: FACE1, FACE2 and FACE3 respectively by using JPEG and JPEG2000 according to compression size for each image.

Table 5: Sequence of "FACE1.bmp" 2D and 3D decompressed image by three methods, according to compressed size

Compressed size	Proposed Method		JPEG2000		JPEG	
	2D RMSE	3D RMSE	2D RMSE	3D RMSE	2D RMSE	3D RMSE
51.6 Kbytes	4.3	0.85	4.0	0.66	4.0	1.8
28.4 Kbytes	4.7	0.82	3.7	1.43	7.6	1.97
16.3 Kbytes	5.4	1.48	4.7	1.8	12.2	116.5
13.5 Kbytes	5.7	1.80	5.1	1.81	NON	NON
11.6 Kbytes	6.1	1.98	5.4	1.93	NON	NON
9 Kbytes	6.7	1.94	5.8	1.86	NON	NON

Table 6: Sequence of "FACE2.bmp" 2D and 3D decompressed image by three methods, according to compressed size

Compressed size	Proposed Method		JPEG2000		JPEG	
	2D RMSE	3D RMSE	2D RMSE	3D RMSE	2D RMSE	3D RMSE
33 Kbytes	2.3	0.55	1.8	0.93	2.7	0.49
16.8 Kbytes	2.6	0.55	2.4	0.97	NON	NON
9.4 Kbytes	3.3	0.59	2.9	0.67	NON	NON
7.7 Kbytes	3.6	0.70	3.2	0.77	NON	NON

VII. Conclusion

This research has presented and demonstrated a new method for image compression used in 3D applications. The method is based on DWT transformation and JPEG transformation with the proposed Minimize-Matrix-Size algorithm. The results showed that our approach introduced better image quality at higher compression ratios than JPEG and JPEG2000 being capable of accurate 3D reconstructing at higher compression ratios. On the other hand, it is more complex than JPEG2000 and JPEG. The most important aspects of the method and their role in providing high quality image with high compression ratios are discussed as follows:

1- Using two transformations, this helped our compression algorithm to increase the number of high-frequency coefficients, and reduces the low-frequency domains leading to increases compression ratios.

2- The Minimized-Matrix-Size algorithm is used to collect each three coefficients from the AC-matrix, to be single floating-point values. This process converts a matrix into an array, leading to increases compression ratios and keeping the quality of the high-frequency coefficients.

3- The PSS-Algorithm represents the core of our parallel search algorithm to finding the exact original data (i.e. decompression algorithm), which is converts a one-dimensional array (i.e. Minimized-Array) to matrix, and depends on the key-values and Limited-Data.

4- The key-values and Limited-Data are used in coding and decoding an image, without these information images cannot be reconstructed.

5- Our approach gives better visual image quality compared to JPEG and JPEG2000. This is because our approach removes most of the block artifacts caused by the 8x8 two-dimensional DCT of the JPEG technique and this is because: Minimize-Matrix-Size algorithm. Also our approach uses single level DWT rather than multi-level DWT in JPEG2000, for this reason blurring removed by our approach.

However, the number of steps of the proposed compression and decompression algorithm more than the JPEG and JPEG2000 steps, also the complexity of PSS-algorithm leads to increased execution time for decompression, because the PSS-Algorithm iterative method is particularly complex.

REFERENCES

[1] A. Al-Haj, (2007) Combined DWT-DCT Digital Image Watermarking, *Science Publications, Journal of Computer Science 3 (9): 740-746,.*

[2] C.Christopoulos, J. Askelof, and M.Larsson (2000) Efficient methods forencoding regions of interest intheupcoming JPEG 2000 still image coding standard,*IEEE Signal Processing Letters*,vol.7,no.9.

[3] G.SadashivappaandK.V.S.AnandaBabu, (2002) PERFORMANCE ANALYSIS OF IMAGE CODING USING WAVELETS, *IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.10.*

[4] I.E. G.Richardson (2002)*VideoCodecDesign*, JohnWiley &Sons.

[5] K. Sayood, (2000) Introduction to Data Compression, 2nd edition, Academic Press, Morgan Kaufman Publishers.

[6] M. Rodrigues, A. Robinson and A. Osman, (2010) Efficient 3D data compression through parameterization of free-form surface patches, In: *Signal Process and Multimedia Applications (SIGMAP), Proceedings of the 2010 International Conference on.* IEEE, 130-135.

[7] M. Rodrigues, A. Osman and A. Robinson, (2013) Partial differential equations for 3D data compression and reconstruction, *Journal Advances in Dynamical Systems and Applications*, accepted for publication, 2013.

[8] M. M. Siddeq, G. Al-Khafaji, (2013) Applied Minimize-Matrix-Size Algorithm on the Transformed images by DCT and DWT used for image Compression, *International Journal of Computer Applications, Vol.70, No. 15.*

[9] M. M. Siddeq(2012)Using Sequential Search Algorithm with Single level Discrete Wavelet Transform for Image Compression (SSA-W), *Journal of Advances in Information Technology. Academic Publisher Vol. 3,No. 4.*

[10] M. M. Siddeq, M. A. Rodrigues(2014) A Novel Image Compression Algorithm for high resolution 3D Reconstruction, *3D Research. Springer Vol. 5 No.2.*DOI 10.1007/s13319-014-0007-6

[11] N. Ahmed, T. Natarajan and K. R. Rao, (1974) Discrete cosine transforms, IEEE Transactions Computer.," vol. C-23, pp. 90-93.

[12] S. Esakkirajan, T. Veerakumar, V. SenthilMurugan, and P. Navaneethan, (2008) Image Compression Using Multiwavelet and Multi-stage Vector Quantization, *International Journal of Signal Processing Vol. 4, No.4, WASET.*

[13] R. C.Gonzalez, R.E.Woods(2001) *Digital ImageProcessing*,AddisonWesley publishing company.

[14] T. Acharya and P. S. Tsai. (2005) *JPEG2000 Standard for Image Compression: Concepts, Algorithms and VLSI Architectures.* New York: John Wiley & Sons.

[15] P. Chen, Jia-Y. Chang (2013) An Adaptive Quantization Scheme for 2-D DWT Coefficients, *International Journal of Applied Science and Engineering Vol.11, No. 1.*

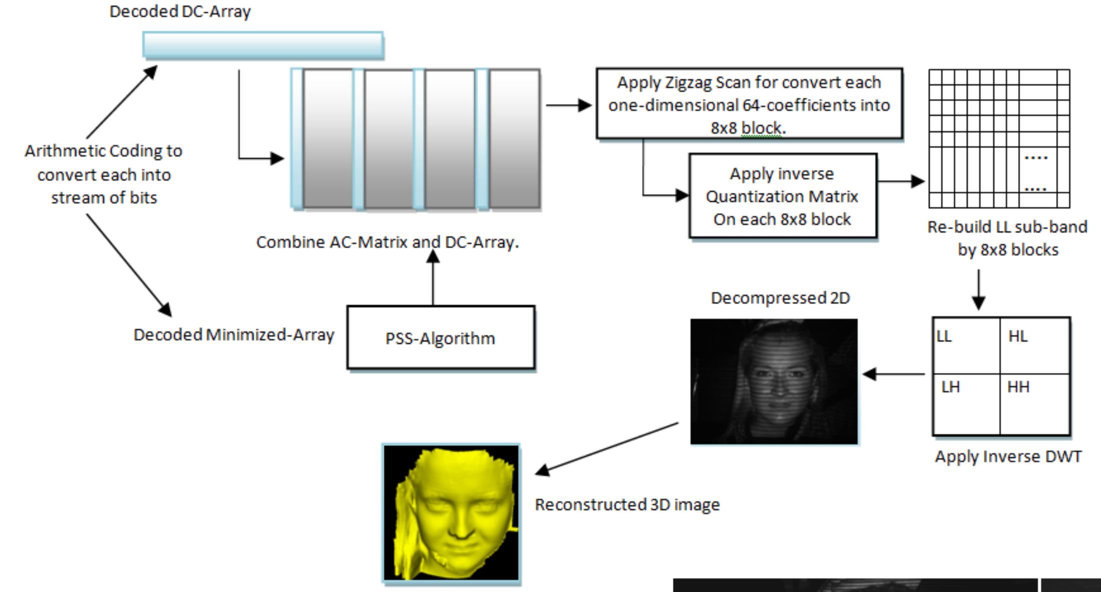
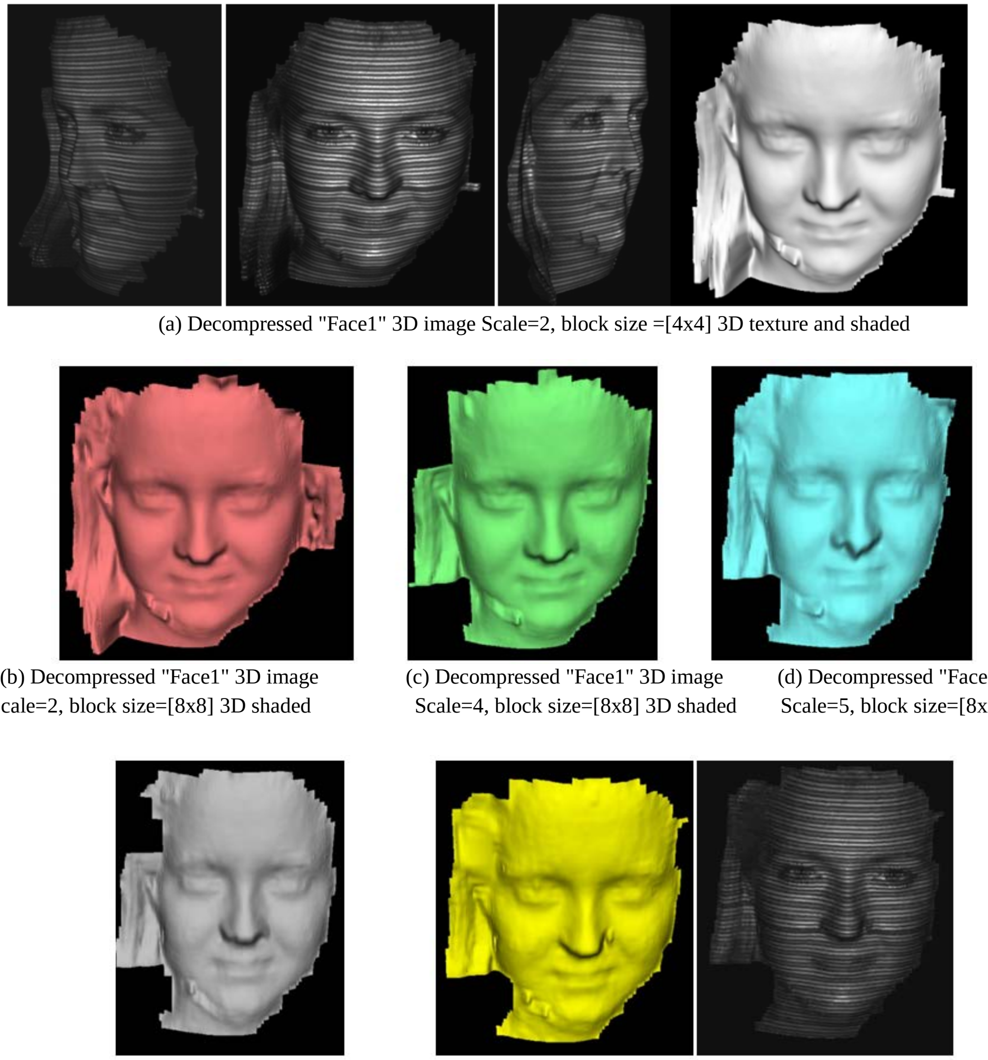

Decoded DC-Array
Arithmetic Coding to convert each into stream of bits
Combine AC-Matrix and DC-Array.
Apply Zigzag Scan for convert each one-dimensional 64-coefficients into 8x8 block.
Apply inverse Quantization Matrix On each 8x8 block
....
....
Re-build LL sub-band by 8x8 blocks
Decoded Minimized-Array
PSS-Algorithm
Decompressed 2D
LL
HL
LH
HH
Apply Inverse DWT
Reconstructed 3D image

Fig-8, flowchart of the proposed Decompression Algorithm

(a) 2D BMP "Face1" (c) 2D BMP "Face3"

Fig-9, (a) and (b) grayscale 2D BMP images, dimension 1392x1040, size=1.38 Mbytes



(a) Decompressed "Face1" 3D image Scale=2, block size =[4x4] 3D texture and shaded

(b) Decompressed "Face1" 3D image Scale=2, block size=[8x8] 3D shaded

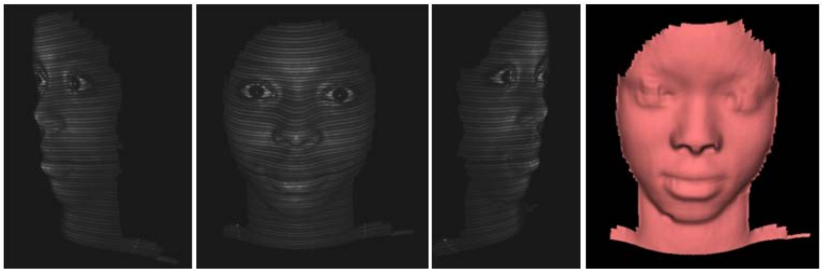
(c) Decompressed "Face1" 3D image Scale=4, block size=[8x8] 3D shaded

(d) Decompressed "Face1" 3D image Scale=5, block size=[8x8] 3D shaded

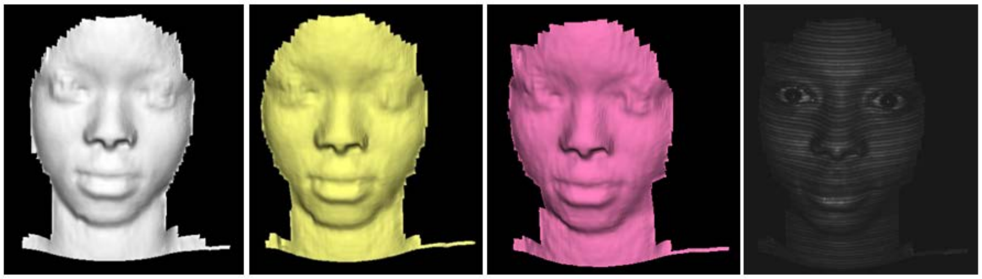
(e) Decompressed "Face1" 3D image Scale=6, block size=[8x8] 3D shaded

(f) Decompressed "Face1" 3D image, Scale=8, block size=[8x8] 3D shaded and texture

Fig-10, (a) Decompressed 3D Face1 image shaded by using high quality parameters applied on the grey-scale image. (b, c, d) Decompressed 3D Face1 image shaded by using normal quality parameters shows the details of 3D surface. (e) Decompressed 3D Face1 image shaded by using low quality parameters and the details of 3D surface still approximately same. (f) Decompressed 3D Face1 image shaded by using very low quality parameters and small amount of the degradation starts appears on the 3D surface.



(a) Decompressed "Face3" 3D image Scale=2, block size =[4x4] 3D texture and shaded

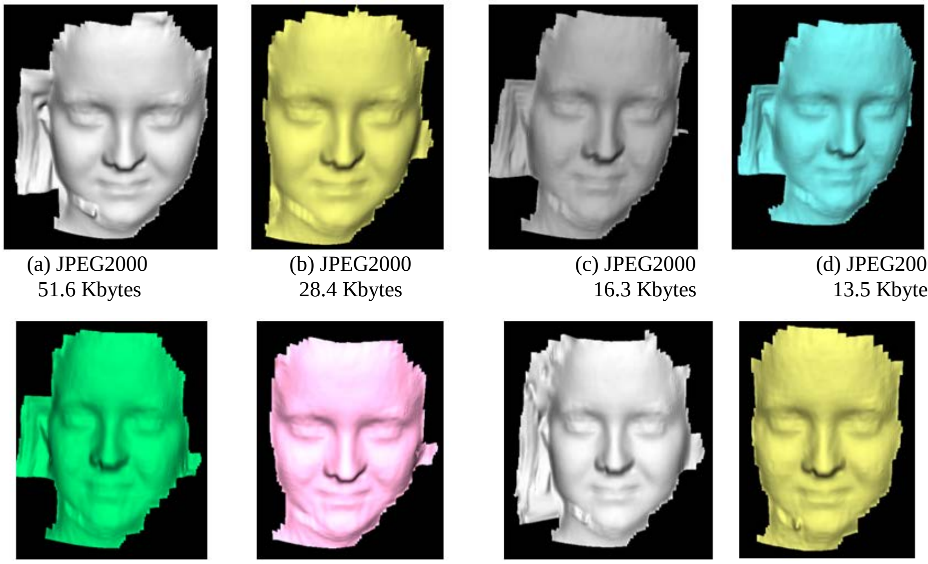


(b) Decompressed "Face3" Scale=2, block size=[8x8]

(c) Decompressed "Face3" Scale=4, block size=[8x8]

(d) Decompressed "Face3" 3D image Scale=5, block size=[8x8] 3D shaded and texture

Fig.-11, (a) Decompressed 3D Face3 image shaded by using high quality parameters applied on the grey-scale image. (b, c) Decompressed 3D Face3 image shaded by using normal quality parameters shows the details of 3D surface is near to original 3D surface. (d) Decompressed 3D Face3 image shaded by using low quality parameters, and small amount of the degradation appeared on the 3D surface.



(a) JPEG2000 51.6 Kbytes

(b) JPEG2000 28.4 Kbytes

(c) JPEG2000 16.3 Kbytes

(d) JPEG2000 13.5 Kbytes

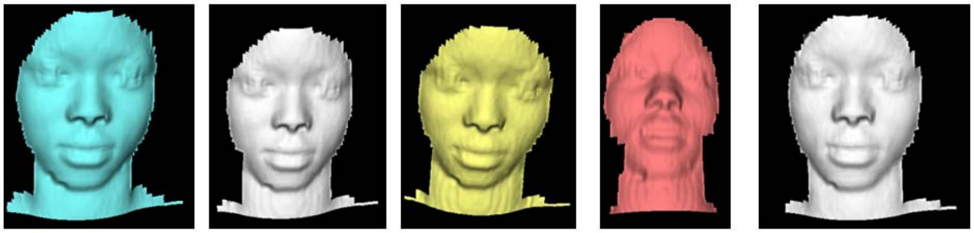
(e) JPEG2000 11.6 Kbytes

(f) JPEG2000 9.0 Kbytes

(g) JPEG 51.6 Kbytes

(h) JPEG 28.4 Kbytes

Fig.-12, (a – f) Decompressed FACE1 image by JPEG2000, (g) and (h) Decompressed FACE1 image by JPEG.



(a) JPEG2000 33 Kbytes

(b) JPEG2000 16.8 Kbytes

(c) JPEG2000 9.4 Kbytes

(d) JPEG2000 7.7 Kbytes

(e) JPEG 33 Kbytes

Fig.-13, (a – d) Decompressed FACE3 image by JPEG2000, (e) Decompressed FACE3 image by JPEG.

National Quality Registries as a Swedish e-Health System

Amra Halilovic, Informatics strategist
County Council of Dalarna
Sweden
amra.halilovic@ltdalarna.se

Abstract - A system of national quality registries (NKRs) has been established in the Swedish health and medical services in the last few decades. Currently, there are 81 NKRs operated with joint financial support from the health authorities and the state. Additional registers are planned, or are under construction. Currently, there are 24 so-called registers candidates who are operated with financial aid. The purpose of this paper is to investigate if NKR can be defined as e-health. An analysis is based on Eyesenbach's definition of "10 e's in e-health". The results of this analysis show that NKRs have four major strengths: Enhancing quality of care, Evidence based, Equity and Ethics. These e's summarise the purpose with 81 NKRs and 24 registers candidates in Sweden. Extended the scope and Enabling information exchange are two strengths, although limited. Patients have to know where to look and how to find information, which is available within Sweden, but not globally.

Keywords— National quality registries, e-health, e-health system.

I. Introduction

The National Quality Registers (NKRs) is a system of quality tools used in Sweden, utilized to continuously improve and provide a good health care. The NKRs contain individualized data concerning patient problems, medical interventions and outcomes after treatment; encompassing all healthcare production in the country. This data is protected by several laws [1].

Currently (2014) there is a trend emerging among administrators, to make register results available to the public over the Internet. Citizens can have direct access to their own care unit quality results, as well as from other care units in the county.

In 2008, Swedish government decided to introduce the health care choices. It means that citizens of any county, when in Sweden, may change care units as often as they want. Health units are not allowed to deny or reject anyone who chooses that unit.

According to recent report from The Swedish National Board of Health and Welfare ("Socialstyrelsen") many citizens find it difficult to choose. There is no readily available information and the choices can be too complex, or absent altogether [2].

NKRs can be used as a solution to this problem. The question is if the usage of NKRs' results would change the definition of NKRs: Is it possible to define NKR as eHealth?

There are more than 51 definitions of e-health [3] - [6] with Eysenbach one being the most quoted [7]:

"e-health is an emerging field in the intersection of medical informatics, public health and business, referring to health services and information delivered or enhanced through the Internet and related technologies. In a broader sense, the term characterizes not only a technical development, but also a state-of-mind, a way of thinking, an attitude, and a commitment for networked, global thinking, to improve health care locally, regionally, and worldwide by using information and communication technology."

II. National Quality Registries

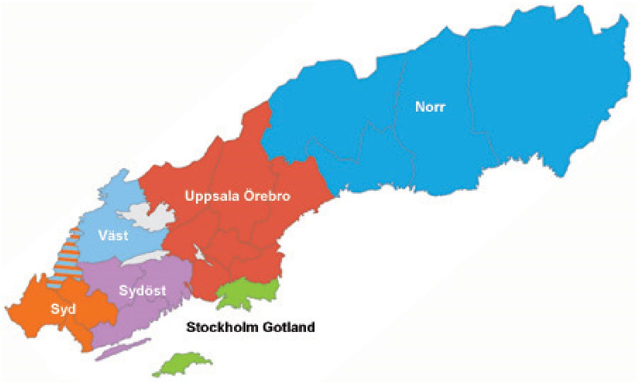
NKRs have a long history. From the beginning, they were created by the individuals who themselves would benefit from them, in their professional lives. NKRs are a system of quality tools with aim to continuously improve and provide good health care. That means that healthcare has to be consistent (equal treatment for all patients, throughout the country) and to ensure that treatments are facts based. NKRs 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.

NKRs also enable:

- Monitoring the progress made in health care, both for the individual patients, as well as the summed group level (for example group of cancer patients).
- Following patient outcomes per county, hospital or a clinic.
- Support the health care work (for example; a checklists).
- Compare healthcare units' own work over time and identity areas of 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).

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.



Figur 1: Six Regional Cancer Centres in Sweden (source: SKL)

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. Strenthen 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 NKRs (approax. 60) have been established [8]: Registercentrum Norr (RCN), Uppsala Clinical Research Center (UCR), QRC Stockholm, Registercentrum Västra Göteland (RVG), Registercentrum SydOst (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 NKRs are available to medical units and county management over the Internet. Citizens do not have access to the results over Internet. Recently, there are some attempts (by some NKRs) to enable reports even to citizens.

Results from NKRs are also accessible by reports like "Open Comparison and Assessment". There reports are freely available to all, including citizens.

III. SWEDEN IN E-HEALTH

Sweden's longstanding commitment to e-health at a regional level was brought together in a national eHealth strategy in 2006. The strategy focuses on:

- Information being able to follow the patient across organizational borders.
- How ICT can help increase patient safety.
- How to increase patient involment in their own care [8].

Authorities are interested in implementation of the National eHealth Strategy and the action plan that the county councils have set up to achieve its targets. This strategy is followed up every year and new goals are set [9].

According to the statistics and several several reports Sweden is at the very forefront in the use of ICT for care documentation and care processen [9], [10].

The characteristic feature of the Swedis model is that a shared health record system is used both for primary care and in-patient care. Electronic Health Record (EHR) and electronic prescriptions have been fully implemented which makes information accessible throught the care chain [9].

Another characteristic feature of Swedis model is collaboration between ICT suppliers, customer groups (county councils) and authorithies such as the Swedish National Board of Health and Welfare, the Swedish Association of Local Authorities and Regions, etc. [9].

The county councils in Sweden, (there are 21 of them), have the responsibility for development and implementation of ICT. They have had different goals and priorities and have made different levels of progress in different areas.

At the moment there are several programs aimed to increase consolidation and coordination of both national services and other services in the county councils.

One example is National Programme for Data collection with its overall goal to develop and implement ICT for automatic data transfer from different EHRs to the ca 100 NKRs.

IV. THE 10 E'S IN "E-HEALTH", ACCORDING TO EYSENBACH

Reference [7] shows that the "e" in e-health does not only stand for "electronic". The author lists ten other "e's", which together best characterize what e-health is about. These are:

1. **Efficiency** – e-health "promises" to increase efficiency in health care by decreasing costs, for example by avoiding duplicative or unnecessary therapeutic interventions through patient involvement.
2. **Enhancing quality** of care – is about improving

quality. E-health may enhance the quality of health care for example by allowing comparisons between different providers.

3. **Evidence based** – all e-health interventions should be evidence-based, for example proven by rigorous scientific evaluation.
4. **Empowerment** of consumers and patients – by access to the knowledge bases of medicine and personal electronic records over the Internet, e-health opens for patient-centred medicine, and enables evidence-based patient choice.
5. **Encouragement** of new relationship between the patient and health professional, for example a true partnership where decisions are made in a shared manner.
6. **Education** of physicians through online sources (continuing medical education) and consumers (health education, tailored preventive information for consumers).
7. **Enabling** information exchange and communication in a standardized way between health care establishments.
8. **Extending** the scope of health care beyond its conventional boundaries. This is meant in both a geographical sense as in conceptual sense. For example e-health enables consumers to easily obtain health services online from global providers.
9. **Ethics** – e-health involves new forms of patient-physician interaction and poses new challenges and threats to ethical issues such as online processional practice, privacy and equity issues.
10. **Equity** – to make health care more equitable is one of the promises of e-health. At the same time we have to be aware of a considerable treat that e-health may deepen the gap between people, for example rich and poor. There have to be political measures which ensure equitable access for all.

V. NKRs as e-Health

Table I (see below) is the summary of characterization of NKRs by the 10 e's in "e-health".

NKRs have four major strengths: **Enhancing** quality of care, **Evidence based**, **Equity** and **Ethics**. These e's summarize the purpose with 81 NKRs and 24 registers candidates in Sweden.

Extended the scope and **Enabling** information exchange are two strengths, although limited. Patients have to know where to look and how to find information, which is available within Sweden, but not globally.

Efficiency, **Empowerment** of consumers and patients and **Education** of physicians and consumers through online sources are potential strengths.

Efficiency can be major strength if it is accessible over Internet. Empowerment of consumers and patients is in the early development stages. Education is possible for physicians, but there are no signs of education for citizens.

Encouragement of new relationship between the patient and health professional is one of NKRs' weaknesses. There are very few evidences that.

Table 1: NKRs as e-health

Nr	The 10 e's in "e-health"	NKRs
1	Efficiency	This can be one of the NKRs' strengths, it is possible to. • Decrease pharmaceutical costs • Find new and efficient medical treatments • Utilize resources
2	Enhancing quality of care	This is major NKRs' strength because it: • Allow comparisons of care units on regional and national level. NKRs enable citizens to choose best care unit. • Allow comparison of best medical methods/treatment - Currently not accessible for citizens. • Allow comparison of best medical artefacts, e.g. prostheses, drugs - Currently not accessible for citizens.
3	Evidence based	• NKRs' data is delivered by patients journal systems. This is major NKRs' strength.
4	Empowerment of consumers and patients	• Not possible today, but by access over the Internet this can be one of the NKRs' strengths.
5	Encouragement of new relationship between patient and health professional	• Not possible today, one weakness of NKRs.
6	Education of physicians and consumers	• This is possible for physicians - by access to the results. Not possible for the patients.
7	Enabling information exchange	• This is one of the NKRs' strengths.
8	Extending the scope	• This is possible in Sweden - not globaly.
9	Ethics	• This is one of the NKRs' strengths.
10	Equity	• This is major NKRs' strength.

VI. Conclusion and Future Work

The ten e's in "e-health" NKRs have:

- Four major strengths
- Two limited strengths
- Three potential strengths
- One weakness

According to these, can we define NKRs as e-health, but the question is if it is meaningful to define one phenomena as e-health if it partially fulfil a list of e's? We invite other views and opinions on this question in the hope that we together can elucidate and delimit the realm of e-health.

References

[1] Nationella kvalitetsregister. Available: http://www.kvalitetsregister.se

[2] *"Valfrihetssystem ur ett befolknings- och patientperspektiv. Slutredovisning."* Socialstyrelsen; 2011. Available: http://www.socialstyrelsen.se

[3] Oh H., Riso C., Enkin M., Jadad A. "What is eHealth? A Systematic Review of Published Definitions" *J Med Internet Res* 2005; 7 (1):el. Available at: http://www.jmir.org.

[4] Della Mea V. "What is e-health (2): the death of telemedicine?" *J Med Internet Res* 2001;3:E22.

[5] Meyers J., Van Brunt D., Patrick K. and Greene, A. (2002). "Personalizing medicine on the Web" *Health Forum Journal, 45*(1), 22-26.

[6] Mitchell, J. *"From telehealth to e-health: The unstoppable rise of e-health"*, Canberra, Australia: Australia: Commonwealth Department of Communications, Information Technology and the Arts (DOCITA); 1999.

[7] Eysenbach, G. "What is e-Health*?"* *Journal of Medical Internet Research*, (3:2), 2001, pp. E20.

[8] Swedish Association of Local Authorities and Regions, http://www.skl.se.

[9] Jelvall L. and Pehrsson T. "eHealth in Swedish County Councils 2012", *Inventory commissioned by the SLIT group*.

[10] European Commission, "eHealth Benchmarking III", Deloitte & Ipsos Belgium, 13 April 2011.

Towards the Flexibility of Software for Computer Network Simulation

Alexander I. Mikov, Elena B. Zamyatina, and Roman A. Mikheev

Abstract**—** This paper discusses network simulator TRIADNS. It is well known that the role of computer networks becomes more important due to progress in new computer technologies (distributed information systems, GRID-computing, Cloud computing and so on). So it is necessary to have effective and flexible program tools for computer network design and simulation. Indeed this program tools have to design computer networks with a lot of computer nodes, so simulation run must be time-consuming. Besides, it is necessary to study topological characteristics of computer network, to investigate traffic, to study computing node's behavior, to test the designed protocols, to study the behavior of routing algorithms and how these algorithms will behave when computer network topology is changed (new nodes can be added to network, some nodes can fail and so on). Network simulator must design and investigate not only hardware, but software too, explore computer networks, considering in particular the specific characteristics of a variety of computer networks.This paper considers approaches allowing to decide these problems: hierarchical model, using ontologies and Data Mining methods for the analyses of simulation results.

***Keywords* -** simulation, computer networks, ontologies, routing algorithms, Data Mining, distributed and parallel simulation.

I. Introduction

Computer networks are very wide spread now. Indeed computer networks are used in distributed information systems, GRID computing, cloud computing and so on.

Widespread computer networks impose requirements to the speed and reliability of information transfer, to its effective treatment. For this reason, it becomes necessary to study traffic, to investigate new protocols, to design and develop new devices and new algorithms.

It is not always possible to apply analytical methods to investigate computer network because of the complexity of modeling object and, moreover, natural experiments can't investigate all aspects of this object too. So the designers prefer to use simulation methods and appropriate program tools (network simulators). A lot of network simulators were developed recently [1]. We consider some of them below.

Because of complexity of modeling object (computer networks) simulators should have the following properties:

- *Simulation experiment should be optimized in respect to time.* Indeed very often it is necessary to investigate large-scale networks with a tremendous amount of computing nodes. It is clear that the simulation of large-scale networks must be terminated within a reasonable time [2, 3]. But it is possible if one can perform simulation experiment on a supercomputer (cluster and so on). Besides, the investigators need the special software tools implementing special synchronization algorithm (conservative or optimistic), managing time advancement [4, 5, 6]. Moreover it is necessary to solve a problem of the equal workload on the computing nodes [7, 8, 9]. And nowadays new class of computer network simulators appears – there are simulators using graphical processors (GPU) [10].
- *A joint study of hardware and software of computer networks.* The computer network designers usually consider separately the hardware and software. However, the most appropriate solution would be to have software tools for design and analysis hardware, design and analysis of algorithms that control hardware, and for the co-design of hardware and software [11]. For example, it is very important to analyze the behavior of routing algorithm after the moment when the topology of computer network is changed (new computing node appears or some nodes become not accessible). In this case, the designer is interested in the topological characteristics of the network. These characteristics may effect the communication complexity of the algorithm. The structure of network may be represented as a graph. So it is important to investigate the structure of network using known graph algorithms (the shortest distance, for example). Nowadays the adaptable routing algorithms are applied in networks. These algorithms change their behavior depending on the values of certain characteristics of the network (overload of communication lines, for example). So it is advisable to simulate routing algorithm. Moreover it is important to simulate the behavior of various devices of computer networks and algorithms which control the behavior of these devices.

[1]This work was supported in part by Russian Foundation for Basic Research (grants 12-07-00302, 13-07-96506).

A. I. Mikov is with the Cuban State University, Krasnodar, Russia (e-mail: Alexander_Mikov@mail.ru)

E. B. Zamyatina is with the National Research University Higher School of Economics, Perm National State Research University, Perm, Russia (e-mail: E_Zamyatina@mail.ru).

R. A. Mikheev is with the Perm National Research State University, Perm, Russia (e-mail: Mikheev@prognoz.ru).

- *Adaptability of software simulators to incorporate into a simulation model new devices and new algorithms that govern their work.* There are various software tools to design the computer networks nowadays. The most popular are: [12] (the design of the local and global networks, multiprocessor and distributed computing systems, the ability to assess the performance of the designed system , etc.); OMNeT + + [13] (a discrete event simulator that allows investigators to explore all levels of computer networks and to include customer modules into simulation model), NS- 2 [14], etc. Each of these simulators has specific characteristics. Some tools are designed to manage local networks, while others permit the design and analyses of global networks. Some of these software tools allow network designing, but have limited modeling capabilities, others are able to perform complex analysis of specific networks (may be only global networks or local or sensor ones). Network simulators have to be able to design, simulate and analyze new types of computer networks, new devices, new algorithms and technologies because of rapid development of network technologies.

The designers and developers of computer networks simulator TriadNS tried to consider the experience of various software tools of this kind. This simulator is based on CAD Triad [15]. The ideas embodied in CAD system Triad allow it to adapt to rapid change of computer networks, new algorithms and technologies due to special linguistic and program tools:

- Linguistic and program tools for the description of the structure of computer networks and the behavior of the devices and computing nodes;
- Advanced analysis subsystem, which includes a library of standard information procedures (information procedures are obtained to collect the information about simulation model during simulation experiment and to process it) and linguistic tools to create new procedures and, therefore, new algorithms of analysis.

Furthermore, the effectiveness of the simulator is provided by distributed (parallel) simulation experiment (using the resources of several nodes of computer network, cluster or multiprocessor (the advantages of a distributed (parallel) simulation experiment are listed in [5, 16]). Optimistic synchronization algorithm (based on knowledge) and load balancing subsystem are implemented in simulator TriadNS. This software permits to reduce the time needed for simulation experiment. Moreover the effectiveness of simulation system may be achieved by the subsystem of collecting and processing of the simulation model characteristics (the processing of data may be partly carried out during simulation experiments) and intelligent analysis of simulation results (based on the methods of Data Mining). Flexibility is achieved through the use of ontologies and the mechanism of redefining models, interoperability (including in the model components developed in the other modeling systems). First of all, we should talk about how the simulation model is presented in the simulator TriadNS, the architecture of simulator and the description of each it's subsystem.

II. The Simulation Model Representation in TriadNS

Simulation model in Triad.Net is represented by several objects functioning according to some scenario and interacting with one another by sending messages. So simulation model is μ={STR, ROUT, MES} and it consists of three layers, where STR is a layer of structures, ROUT – a layer of routines and MES – a layer of messages appropriately. The layer of structure is dedicated to describe objects and their interconnections, but the layer of routines presents their behavior. Each object can send a message to another object. So, each object has the input and output poles (Pin – input poles are used to send the messages, Pout – output poles serve to receive the messages). One level of the structure is presented by graph P = {U, V, W}. P-graph is named as graph with poles. A set of nodes V presents a set of programming objects, W – a set of connections between them, U – a set of external poles. The internal poles are used for information exchange within the same structure level; in contrast, the set of external poles serves to send messages to the objects situated on higher or underlying levels of description. Special statement <message> through <name of pole> is used to send the messages.

One can describe the structure of a system to be simulated using such a linguistic construction:

structure <name of structure> ***def*** (<a list of generic parameters>) (<a list of input and output parameters>) <a list of variables description> <statements>) ***endstr***

Special algorithms (named "routine") define the behavior of an object. It is associated with particular node of graph P = {U, V, W}. Each routine is specified by a set of events (E-set), the linearly ordered set of time moments (T-set), and a set of states {Q-set}. State is specified by the local variable values. Local variables are defined in routine. The state is changed if an event occurs only. One event schedules another event. Routine (as an object) has input and output poles (Pr_{in} and Pr_{out}). An input pole serves to receive messages, output – to send them. One can pick out input event e_{in}. All the input poles are processed by an input event, an output poles – by the other (usual) event.

routine<name>(<a list of generic parameters>)(<a list of input an0d output formal parameters>) ***initial*** <a sequence of a statements> endi event <a sequence of a statements> ***ende*** ***event*** <a name of an event> <a sequence of statements> ***ende*** … ***event***<a name of an event><a sequence of a statements> ende ***endrout***

The investigator may not describe all the layers. So if it is necessary to study structural characteristics of the model, only the layer of structures can be described. The example of computer network (the layer of structure) is given below. This computer network consists of a server and several clients.

Structure Client_Server[integer Number_of_Clients] ***def***
Client_Server := ***node*** Server<Receive, Send> +

```
        node Клиент[ 0 : Number_of_Clients - 1 ] < Receive,
        Send >;
         integer i;
         for i := 0 by 1 to Number_of_Clients  - 1 do
             Number_of_Clients := Number_of_Clients  +
              arc ( Client[ i ].Send -- Сервер.Receive ) +
               arc ( Сервер.Send -- Клиент[ i ].Receive );
         endf;
    endstr
```

Fig.1. The Structure of Computer Network Client_Server

Note, please, that the layer of structure is a procedure with parameters. Triad-model is considered as a variable. Initially it may be void and further may be constructed with the special statements of Triad-language (operations within the layer of structures.

Fig.1. gives the structure of network "Client_Server". It consists of the node "Server" and the attached array of nodes "Client". The links between nodes are set within the cycle for with the help of arcs. Input and output poles have to be specified: (arc (Server.Send -- Client[i].Receive)). The number of nodes Client may be changed by formal parameter Number_of_Client.

Client behavior scenario is described with special linguistic unit routine. The syntax of routine is given above. The description of the "Client" behavior is given below:

```
routine  Client ( input Receive; output  Send )[ real deltaT ]
   initial boolean Quiery_is_Send;
        Quiery_is_Send := false; schedule  Quiery in 0;
       Print "Client Initialization";
   endi
   event Quiry; (* it is an event *)
        out "I send a quiry" through Send;
        Print "A Client sends a quiry to Server";
       schedule ЗАПРОС in deltaT;
   ende
   endrout
```

Fig.2. The Routine "Client"

The routine is a procedure with parameters too, it includes not only the interface parameters (input and output interface parameters "Receive" and "Send", but the parameter deltaT-the time interval between the queries of Clients to Server.

The instances of routine are formed by the statement ***let Client*** (clientDeltaT) ***be*** Client. An instance of routine may be "put" on an appropriate node with the help of statement: ***put*** Client ***on*** Model.Client[i]<Receive=Receive,Send=Send>. The input and output poles of routine are matched to the poles of node here.

There are two ways to describe model in Triad: via text editor or via graphical editor. The description of a layer of structure being built with the help of graphical editor is given below. This description is a fragment of computer network. It consists of several workstations sending messages between them. Besides, the computer network includes the routers responsible for the searching of the route.

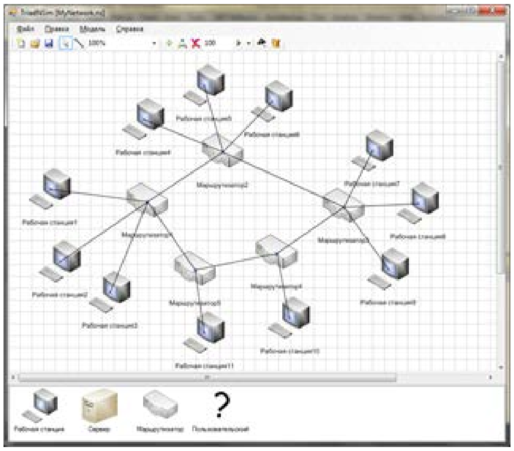


Fig.3. The fragment of computer network. Graphical editor.

The description of this fragment of computer network being built with the help of text editor is given on fig.4.

```
Type Router,Host; integer i;
M:=dcycle (Rout[5]<Pol>[5]);
M:=M+node (Hst[11]<Pol>);
for i:=1 by 1 to 5 do
          M.Rout[i]=>Router;
      M:=M+edge(Rout[i].Pol[1] — Hst[i]);
endf
for i:=1 by 1 to 3 do
      M:=M+edge(Rout[i].Pol[2] — Hst[2*i-1]);
endf;
for i:=0 by 1 to 11 do M.Hst[i]=>Host; endf;
```

Fig.4.The fragment of computer network in Triad language

Simulation model (see fig.4.) is built using graph constants. A set of special linguistic units - graph constants - presents the basic types of topologies of computer network. In the text given above the graph constant "directed cycle" (Dcycle) was used. Besides, in above example the semantic types (Type Router,Host) were used. Namely they are "router" and "host". The semantic types are used for simulation model redefining. More details will be given later.

There are the several standard procedures in the structure layer. The investigator is able to take out from the structure of model a lot of characteristics: a set of nodes, a set of arcs, a set of edges and etc. Moreover one can find the shortest distance between two nodes, connected components (procedure GetStronglyConnectedComponents(G)), selection of the structure layer (procedure GetGraphWithoutRoutines(M)) and so on.

Besides, the investigator obtains the linguistic and programming tools enabling him to write the absent procedure by himself.

The investigation of the structure layer only is static process. The simulation process may take place only after the definition of the behavior of all nodes. As it was noted above the behavior is determined by the statement ***Put***.

It is well known that a simulation is a set of object functioning according to some definite scenarios controlled by synchronizing algorithm. The simulation run is initialized by the statement ***simulate***:

Simulate <список моделей> ***on condition of simulation*** <имя условия моделирования>(<настроечные параметры>)(<параметры интерфейса>)(<список информационных процедур>; <последовательность операторов>).

One can pay an attention to the fact that the several models may be simulated under the same conditions of simulation simultaneously.

III. The Algorithm of Investigation

The objects of simulation model are managed by the special algorithm during the simulation run. Let us name it as "simulation algorithm" (CAD system Triad has distributed version and corresponding algorithm for distributed objects of simulation model too) [15]. CAD system Triad includes analyses subsystem implementing the algorithm of investigation - special algorithm for data (the results of simulation run) collection and processing.

The analysis subsystem includes special objects of two types**:** *information procedures* and *conditions of simulation*. Information procedures are "connected" to nodes or, more precisely, to routines, which describe the behavior of particular nodes during simulation experiment. Information procedures inspect the execution process and play a role of monitors of test desk.

Conditions of simulation are special linguistic constructions defining the algorithm of investigation because the corresponding linguistic construction includes a list of information procedures which are necessary for investigator.

The algorithm of investigation is detached from the simulation model. Hence it is possible to change the algorithm of investigation if investigator would be interested in the other specifications of simulation model. For this one need to change the conditions of simulation. But the simulation model remains invariant. We may remind that it is not possible in some simulation systems.

One can describe the information procedure as so:

information procedure*<name>(<a list of generic parameters>)(<input and output formal parameters>)*

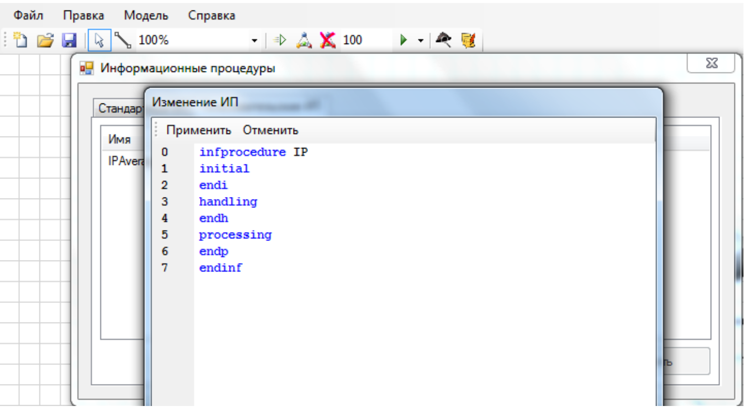
initial *<a sequence of statements>* ***endi***

*<a sequence of statements>****processing*** *<a sequence of statements>…****endinf***

It is possible to examine the value of local variables, the event occurrence and the value of messages which were sent or received. A part of linguistic construction 'processing' defines the final processing of data being collected during simulation run (mean, variance and so on). Let us present the linguistic construction ***conditions of simulation***:

Conditions of simulation*<name>*(<a list of generic parameters>)(<input and output formal parameters>) **initial** <a sequence of statements> ***endi*** <a list of information procedures> <a sequence of statements> ***processing*** <a sequence of statements>…***endcond***

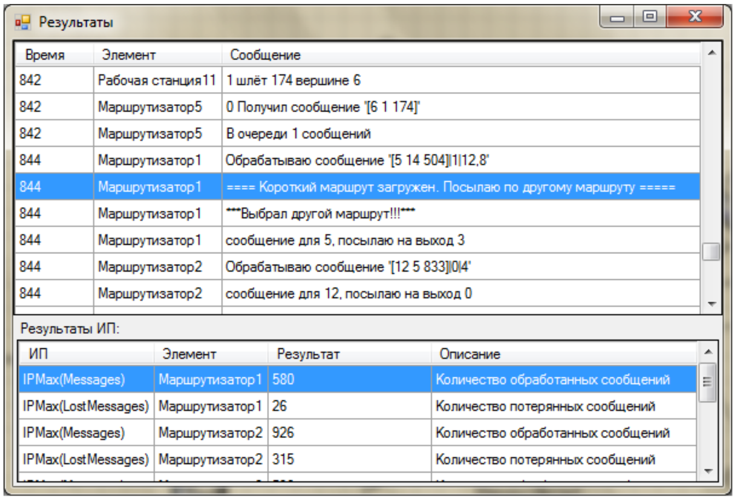
The linguistic construction ***conditions of simulation*** describes the algorithm of investigation which defines not only the list of information procedures but the final processing of some information procedure and checks if conditions of simulation correspond to the end of simulation.



```
infprocedure IP
initial
endi
handling
endh
processing
endp
endinf
```

Fig.5. The form for information procedure

The subsystem of visualization represents the results of simulation. One can see the representation of the results of simulation run at fig.6.



Время	Элемент	Сообщение
842	Рабочая станция11	1 шлёт 174 вершине 6
842	Маршрутизатор5	0 Получил сообщение '[6 1 174]'
842	Маршрутизатор5	В очереди 1 сообщений
844	Маршрутизатор1	Обрабатываю сообщение '[5 14 504]|1|12,8'
844	Маршрутизатор1	==== Короткий маршрут загружен. Посылаю по другому маршруту =====
844	Маршрутизатор1	***Выбрал другой маршрут!!!***
844	Маршрутизатор1	сообщение для 5, посылаю на выход 3
844	Маршрутизатор2	Обрабатываю сообщение '[12 5 833]|0|4'
844	Маршрутизатор2	сообщение для 12, посылаю на выход 0

Результаты ИП:

ИП	Элемент	Результат	Описание
IPMax(Messages)	Маршрутизатор1	580	Количество обработанных сообщений
IPMax(LostMessages)	Маршрутизатор1	26	Количество потерянных сообщений
IPMax(Messages)	Маршрутизатор2	926	Количество обработанных сообщений
IPMax(LostMessages)	Маршрутизатор2	315	Количество потерянных сообщений

Fig.6. The results of simulation

IV. The Components of Simulation System TriadNS

Let us consider simulation modeling system TriadNS, its appointment, its components and functions of each component. So simulation system Triad.Net is a modern version of previous simulation modeling system Triad [6] dedicated to computer aided design and simulation of computer systems. Triad.Net is designed as distributed simulation system, so various objects of simulation model may be distributed on the different compute nodes of a computer system. One more specific characteristic of Triad.Net – remote access, so several investigators may fulfill a certain project from different computers situating in different geographical points.

Distributed simulation system Triad.Net consists of some subsystems: compiler (TriadCompile), core of simulation system (TriadCore), graphical and text editors, subsystem of testing and debugging (TriadDebugger), subsystem of distributed simulation (synchronization of simulation model objects which are situated on different compute nodes of computer system, conservative and optimistic algorithms realization)(TriadRule), subsystem for equal workload of compute nodes (TriadBalance), subsystem of remote and local access (TriadEditor), subsystem of automatic and semiautomatic simulation model completeness (TriadBuilder), the subsystem for remote access and a security subsystem from external and internal threats TriadSecurity), the subsystem of automatically extending the definition of the model (TriadBuilder), the subsystem of intellectual processing of the results of simulation experiment (TriadMining). Initially we address to the specific characteristics of simulation model in TriadNS.

V. Knowledge Representation

It is important to involve into the simulation process not only the specialists in simulation but the specialist in specific domains and specialists in the other spheres of knowledge. That is why it is necessary to adjust a simulation system to specific domain. Indeed the investigator of computer network may use a graph theory while studying the structure of network, or a queue network theory, or the theory of Petri Nets. Ontologies are used in TriadNS to adjust the simulation system to specific domain.

Ontologies can be applied on the different stages of simulation [17, 18]. Very often ontologies are applied for the simulation model assembly. So the simulation model may consist of separately designed and reusable components. These components may be kept in repositories or may be found via Internet. The ontologies keep the information about interconnections of simulation model components and other characteristics of these components.

Ontologies enable investigators to use one and the same terminology.

Ontologies allow to make the repositories of components to store not only an information about their characteristics, interfaces, but the information about their interconnections.

The base ontology is designed in TriadNS.

Its basic classes are: TriadEntity (any named logic entity), Model (simulation model), ModelElement (a part of simulation model and all the specific characteristics of a node of structure layer), Routine (node behavior), Message (note, please, that structure layer nodes of simulation model can interchange with messages) and so on.

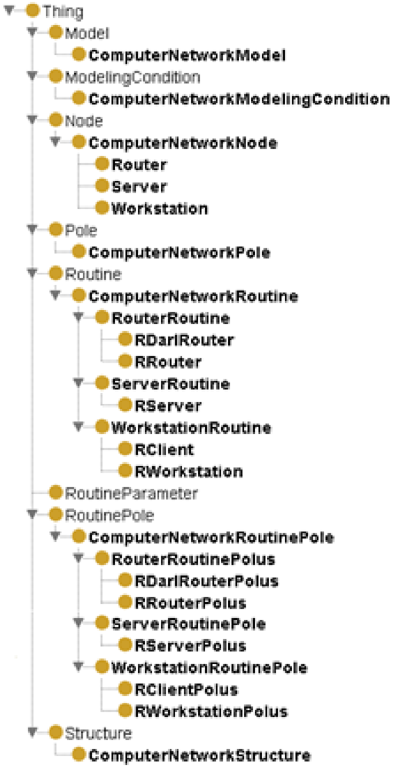


Fig. 5. The hierarchy of the classes of the ontologies in TriadNS

The basic properties of base ontology are:

- *The property of ownership:* model has a structure, a structure has a node, a node has a pole and so on.
- *The property to belong to something* –- inverse properties to previous one– The structure belongs to the model, the node belong to strucrure, the pole belong to the node and so on.
- *The properties of a pole and an arc connection:* connectsWithArc (Pole, Arc), connectsWithPole (Arc, Pole).
- *The property of a node and an appropriate routine binding*-putsOn (Routine, Node).
- *The properties of a node and an appropriate structure binding:* explicatesNode (Structure, Node), explicatedByStructure (Node, Structure).
- *The property of the model and conditions of simulation binding* (Model, ModelingCondition).

The simulator TriadNS has some additional special subclasses of the base classes (specific domain – computer networks). (fig.6.):

- *ComputerNetworkModel* (a model of a computer network), ComputerNetworkStructure (a structure of a computer network model).
- *ComputerNetworkNode* (a computer network element, it contain several subclasses: Workstation, Server, Router).
- *ComputerNetworkRoutine* (a routine of a computer network) и т.д.

This ontology includes two special properties of a pole. These properties are used to *check* the conditions of matching routine to a node:

- *isRequired(ComputerNetworkRoutinePole, Boolean)* – this property check if it is necessary to connect a pole with another pole?
- *canConnectedWith(ComputerNetworkRoutinePole, ComputerNetworkRoutine)* –this property check the semantic type of an element of a structure being connected.

VI. Redefining of simulation model

An ordinary simulation system is able to perform a simulation run for a completely described model only. At the initial stage of designing process an investigator may describe a model only partly omitting description of behavior of a model element μ_{r*} = {STR, ROUT*, MES}). Simulation model may be described without any indication on the information flows effecting the model (μ_{s*} = {STR*, ROUT*, MES}) or without the rules of signal transformation in the layer of messages (μ_{m*} = {STR, ROUT*, MES}). However for the simulation run and the following analysis of the model all these elements have to be described may be approximately.

For example, in a completely described model each terminal node $v_i \in V$ has an elementary routine $r_i \in$ ROUT. An elementary routine is represented by a procedure. This procedure has to be called if one of poles of node v_i receives a message. But some of the terminal nodes v_i of partly described model do not have any routines. Therefore the task of an automatic completion of a simulation model consists either in “calculation” of appropriate elementary routines for these nodes, i.e. in defining $r_i = f(v_i)$, either in “calculation” of a structure graph $s_i = h(v_i)$ to open it with (in order to receive more detailed description of object being designed). It was mentioned above that the routine specifies behavioral function assigned to the node, but the structure graph specifies additional structure level of the model description. And at the same time, all structures s_i must be completely described as the submodels.

These actions have to be fulfilled by the subsystem TriadBuilder.

Subsystem TriadBuilder [19] attempts to search the appropriate routine by the help of base ontology (it was described earlier). It may be found thanks to special semantic type (semantic type “Router” and “Host”, for example).

Model completion subsystem starts when the internal form of simulation model is built according to a Triad code.

First, *model analyzer* searches the model for incomplete nodes, and marks them. Thus, the model analyzer will mark all *Rout* nodes. After the inference module starts looking for an appropriate routine instance for each of marked nodes according to specification condition (the semantic type of node and routine must coincide).Then the condition of configuration must be checked (the number of input and output poles of node and the number of poles of routine must coincide).

After the appropriate instance has been found, it may be put on the node.

VII. Intellectual analysis of the simulation experiment results

It is well known that the goal of a simulation experiment is to obtain the most accurate and adequate characteristic of the studied object. This stage of simulation deals with data collection and processing. The special syntax units such as information procedures and conditions of simulation are designed in TriadNS. Information procedures and conditions of simulation are described above. Note, please, that data collection and data processing with the help of information procedures permit to obtain more adequacy results. Information procedures monitor only these characteristics of simulation model which are interested for investigator. In contrary some other simulators able to monitor and to collect a set of predefined characteristics.

But we can note another problem: the results of simulation experiment are not ordered and not structured. The processing of a simulation experiment results requires highly skilled analysts.

So we can state the appearance of several papers with the suggestion to make the additional processing of the results of simulation experiments [20] and to apply the methods of Data Mining for these purposes [21].

Usually investigators obtain standard report with the results of simulation. The additional processing allow to find dependences between characteristics of the modelling objects.

The analyses of these dependences allow to reduce the overall data capacity, dimension of problem and eventually to optimize the simulation experiment.

The additional processing may be done with the special software tools of TriadNS (component TriadMining). TriadMining use the results of the information procedures, the results are processed with the help of regression analyses, time serious, Bayesian networks and so on. We mentioned above that an information procedure monitors the implementation of the sequence of events, the variables changing and so on. It is well known that the sequence of the predefined events allow to find crashes in nodes of telecommunication systems. Here is an example of information procedure.

```
information procedure event_sequence (in ref event
E1,E2,E3;out Boolean arrived)
initial interlock (E2,E3); Arrived := false;
  case of e1:available(e2);
     e2:available( E3):
              e3:ARRIVED:=true;
  endc
endinf
```

Fig. 6. The information procedure to detect the proper sequence of events

So investigator may detect the arrival of the sequence of events E1→E2→E3. The statement ***interlock*** provides input parameter blocking (event E1 in this case). It means that information procedure doesn’t watch parameters being marked in interlock statement. The statement ***available*** allows beginning the marked parameter monitoring again.

Information procedure monitors the changing of variables and the moments of appropriate time. So the time series may be formed. It is necessary to analyze the similarity of two or more time series. So it is possible to find dependences between the elements of simulation model and reduce the data capacity.

VIII. CONCLUSION

The paper discusses the problems of flexible software for computer network simulation. Authors consider ontology approach application to automatic redefining of simulation model and to adjusting the simulation system to the specific domain.

Simulator TriadNS is provided with a convenient graphical interface. Simulator permits separate and joint hardware and software modelling. Another distinguished characteristic of the simulator is the ability to make a distributed simulation experiment.

The Data Mining methods allow to simplify the analyses of the simulation experiment results. Ontologies enable to automate the simulation model construction and to achieve the interoperability of the software tools (to use components designed in the other simulation systems).

REFERENCES

[1] S. Salmon, H.ElAarag. Simulation Based Experiments Using Ednas: The Event-Driven Network Architecture Simulator. In Proceedings of the 2011 Winter Simulation Conference S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds. The 2011 Winter Simulation Conference 11-14 December 2011. Grand Arizona Resort Phoenix, AZ, pp. 3266-3277.

[2] A.I.Mikov, E.B.Zamyatina The simulation model technologies for big systems investigation // In Proceedings of the Scientific Conference "Scientific service on the Internet" – M.: MSU, 2008. C.199-204.[in Russian]

[3] Y.Liu, Y.He. A Large-Scale Real-Time Network Simulation Study Using Prime. M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds. The 2009 Winter Simulation Conference 13-16 December 2009. Hilton Austin Hotel, Austin, TX, pp. 797-806.

[4] Riley, R.M. Fujimoto, M. Ammar. A Generic Framework for Parallelization of Network Simulations”, in Proc. 7th Int.Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 1999, p. 128-135.

[5] R.M. Fujimoto Distributed Simulation Systems. In Proceedings of the 2003 Winter Simulation Conference S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds. The 2003 Winter Simulation Conference 7-10 December 2003. The Fairmont New Orleans, New Orleans, LA, pp. 124-134

[6] E. Zamyatina, S. Ermakov. The Synchronization Algorithm of Distributed Simulation Model in TRIAD.Net. Applicable Information Models. ITHEA, Sofia, Bulgaria, 2011, ISBN: 978-954-16-0050-4, pp.211-220.[in Russian]

[7] L. F. Wilson, W. Shen Experiments in load migration and dynamic load balancing in Speedes // Proc. of the Winter simulation conf. / Ed. by D. J. Medeiros, E. F.Watson, J. S. Carson, M. S. Manivannan. Piscataway (New Jersey): Inst. of Electric. and Electron. Engrs, 1998. P. 487–490.

[8] G. Zheng Achieving high performance on extremely large parallel machines: Performance prediction and load balancing: Ph.D. Thesis. Department Comput. Sci., Univ. of Illinois at Urbana-Champaign, 2005. 165 p. [Electron. resource]. http://charm.cs.uiuc.edu/.

[9] A.I.Mikov, E.B.Zamyatina, A.A.Kozlov The Multiagent Approach to the Equel Distribution of the Workload. Natural and Artificial Intelligence, ITHEA, Sofia, Bulgaria, 2010, pp.173-180.

[10] L. Djinevski., S. Filiposka, D.Trajanov Network Simulator Tools and GPU Parallel Systems. In Proceedings of Small Systems Simulation Symposium 2012, Niš, Serbia, 12th-14th February 2012, pp.111-114

[11] W.Hu, H.S. Sarjoughian A Co-Design Modeling Approach For Computer Network Systems. . In Proceedings of the 2007 Winter Simulation Conference S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds. The 2007 Winter Simulation Conference 9-12 December 2007 J.W. Marriott Hotel, Washington, D.C., pp. 124-134

[12] [NS-2. 2004] The Network Simulator - NS-2. Доступно на сайте: http://www.isi.edu/nsnam/ns [Проверено 21 марта 2012]

[13] [OPNET, 2004] OPNET Modeler. Доступно на сайте: <http://www.opnet.com> [Проверено: 21 марта 2012]

[14] [OMNeT++, 2005] OMNeT++ Community Site. Доступно на сайте: http://www.omnetpp.org. [Проверено: 21 марта 2012]

[15] A.I. Mikov Simulation and Design of Hardware and Software with Triad// Proc.2nd Intl.Conf. on Electronic Hardware Description Languages, Las Vegas, USA, 1995. pp. 15-20.

[16] R.E. Nance Distributed Simulation With Federated Models: Expectations, Realizations And Limitations. In Proceedings of the 1999 Winter Simulation Conference. P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds., The 1999 Winter Simulation Conference 5 – 8 December 1999 Squaw Peak, Phoenix, AZ, pp. 1026-1031.

[17] P Benjamin., K.V Akella., K Malek., R Fernandes. An Ontology-Driven Framework for Process-Oriented Applications // Proceedings of the 2005 Winter Simulation Conference / M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.,– pp 2355-2363

[18] P. Benjamin.,M. Patki, R. J Mayer. Using Ontologies For Simulation Modeling // Proceedings of the 2006 Winter Simulation Conference/ L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds. –pp.1161-1167

[19] A.Mikov A., E.Zamyatina, E. Kubrak. Implementation of simulation process under incomplete knowledge using domain ontology. In proceedings of 6-th EUROSIM Congress on modeling and Simulation. 9-14, September, 2007, Ljubljana, Slovenia, Vol.2. Full papers, 7 pp.

[20] G. Neumann, J.Tolujew , From Tracefile Analysis to Understanding the Message of Simulation Results, proceeding of the 7th EUROSIM Congress on Modeling and Simulation, Prague, Czechia, 2010, 7 pp.100-117

[21] T. Brady, E.Yellig, Simulation Data Mining: a new form of simulation output, 37th Winter Simulation Conference, Orlando, USA, 2005, pp 285-289.

Future Internet Architecture: the Connected Device Interface

Pierangelo GARINO[1], Letterio ZUCCARO[2], Guido ODDI[2], Andi PALO[2], Andrea SIMEONI[2]

Abstract— unsupervised widespread of telecommunications, multimedia devices, Internet technologies, services and contents lead to the "fragmentation problem". To build up an effective Future Internet it is essential to face the current Internet limitations. Several solutions to face the Internet heterogeneity exist, but none of them proposes a holistic and standardizable approach to interface the connected devices to the network. In the present work we present a Future Internet enabler called Connected Device Interface (CDI), that is in charge of addressing the fragmentation problem, proposing a unified and extended set of APIs counting on both best available standard solutions as well as completely novel functionalities. The CDI architecture is analysed in detail, its API set is described and an instance implementation is presented.

Keywords— Future Internet, Generic Enabler, Fragmentation problem, Connected Device, Interfaces

I. Introduction

It is a fact that the Internet revolution began in nineties and is still leading the technology evolution, indeed most of the current strategic trends rely on it ([1], [16]). It is a tangible fact that the amount of available services and contents grows as well as the ways they can be accessed. Internet has become in few years an open business arena where content and service providers, network providers and device manufacturers fight to gain the market share. But the freedom of the Internet-based business has its own drawbacks. Driven by innovation and offer differentiation, device vendors and manufacturers are continuously introducing new features and capabilities in their products. Such vertical businesses introduce severe interoperability and portability limitations, both due to the lack of standards and to the existence of heterogeneous technologies (e.g., native programming, scripting, codecs, platforms, app markets). This is known as the "fragmentation problem" because it poses high barriers to the development of innovative future network applications, to the remote management of device connectivity functionalities, but it also decreases network applications ecosystem revenues and increases expenses and network and mobile energy consumption (e,g, see [11], [13] and [21]) . The only constraint in a similar scenario, is that users are willing to access their favourite services and contents by means of devices connected to the Internet through access technologies managed by a limited set of network operators. Such consideration inspired some edge research works [2] as well as standardization bodies and industrial initiatives worldwide to redesign the Future Internet fundamentals with the aim to address the fragmentation problem.

In compliance with the Future Internet architecture described in [2], [3] and [20], where the Generic Enabler (GE) concept is introduced, in the present work we address the definition of a GE implementing a common set of interfaces to the connected devices aiming to extend and influence the existing standards and initiatives (e.g., W3C, GSMA, PhoneGap) active in the field. The GE is called Connected Device Interface (CDI) and it is part of a more comprehensive Future Internet Core Platform developed within the pan European FI-WARE project. It is worth noticing that GEs are built on top of existing communication network technologies: the control commands generated by the GEs, in fact, are then enforced by the underlying networks by means of their specific resource management procedure (e.g., admission control [5], [6,15], routing [7], load balancing [8,18], congestion control [9], scheduling [10], resource discovery [14] and allocation [20], quality of experience [4,22] and security [17]). In section 2 we analyse the state of the art solutions available to overcome the fragmentation problem, highlighting their limitations that call for a novel solution. In section 3 we describe the CDI architecture from a functional point of view, together with details about the offered APIs. An instance implementation of the CDI GE is presented in section 4. A brief analysis of the potential business benefits is done in section 5. Some considerations and future works are reported in the conclusive section.

II. State of the Art

It is recognised that the fragmentation problem is caused by the coexistence of dissimilar software and hardware platforms (e.g. Grid [12]) adopted for connected devices, and by the variety of incompatible OSes and programming languages. This heterogeneity is introducing several troubles to develop once for all applications and to make them run easily on all such platforms. Moreover, the lack of standard interfaces to control and manage broad ranges of different devices poses high barriers for applications innovation and interoperability.

The early tentative to fill the gap has been the adoption of middleware based technologies (Java, Flash, Shockwave, just to mention the most known), given their ability to abstract native platforms, boosted the portability of applications. However the adoption of the very same middleware has not been standardized, and thus cannot be imposed to such a broad and fragmented market. Taking this principle in mind, we

[1] Author is with Telecom Italia S.p.A., Via G. Reiss Romoli 274, Turin, 10148, Italy (email: pierangelo.garino@telecomitalia.it)

[2] Authors are with the Department of Computer, Control and Management Engineering "A. Ruberti" at "Sapienza" University of Rome, Via Ariosto 25, 00185 Rome, Italy (e-mail: surname@diag.uniroma1.it).

decided to exploit the natural convergence of native applications toward web based technologies. Indeed most of the existing platforms already support, or at least are going to support, web based environments (browsers, web runtimes) and languages (HTML-HTML5, CSS and JavaScript). This broad and implicitly agreed support for web technologies, makes it the least common denominator on top of which any convergent solution to the fragmentation problem can rely on. We strongly believe that Future Internet applications will be based on web standards. This consideration is supported by the fact that standardization bodies and industrial initiatives work to define common sets of interoperable web based APIs, giving to Future Internet applications full access to native device capabilities and enhanced communication means with the external world.

For what concerns the interaction between the application and the device capabilities the main standardization effort is carried on by the W3C, aimed to produce detailed WebIDL API specifications, to be implemented and supported by most of the web browsers and web runtimes. For what concerns the remote device configuration the main initiative is carried out by the Open Mobile Alliance Device Management working group. It specifies protocols and mechanisms to perform management of devices by the exposure of RESTful APIs. The CDI aims to select the best set of APIs, extend them to fill the gap and to fuse them into a unique, homogeneous and standardizable GE representing a milestone in the field of interfacing the connected device to the network.

III. CDI ARCHITECTURE

The CDI is in charge of providing to Future Internet applications and services the possibility to exploit the device features and capabilities, through the implementation of interfaces and related Application Program Interfaces (APIs) towards the connected devices. With the term “connected devices” we refer to a broad range of networked electronic components (e.g. mobile phones, tablets, Smart TVs, Set-Top-Boxes, In-Vehicle systems, etc.) each being able to connect to at least a communication network. As shown in Figure 1, CDI provides three sets of APIs, identified by the on-device, off-device and Mobility Manager interfaces.

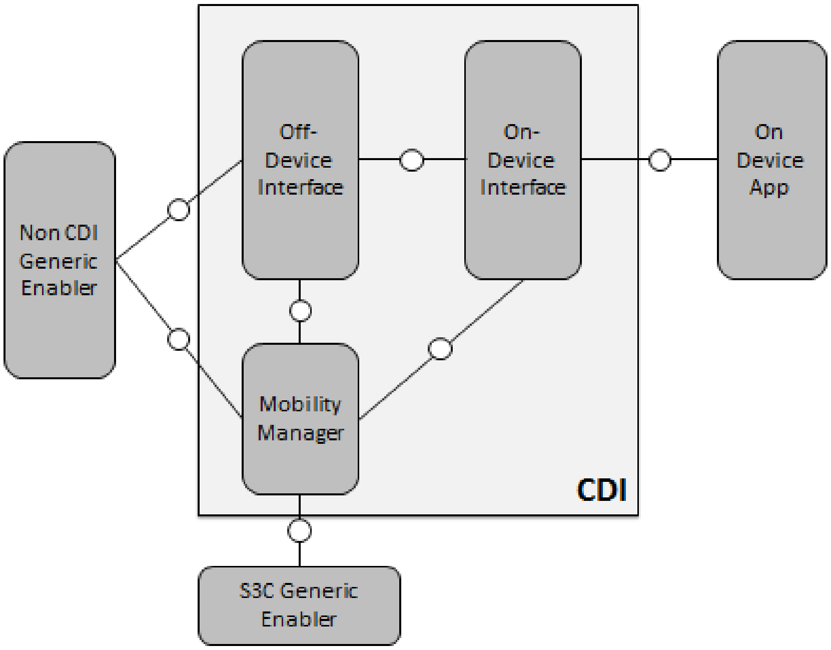


Figure 1 - High level description of the CDI’s architecture in FMC (Fundamental Modelling Concepts)

On-device APIs offer functionalities to local running applications, and are aimed to bridge the portability gap between many different platforms, by exposing a common set of APIs to exploit native devices’ capabilities. This subsystem contains also models and primitives for the assessment of the current Quality of Experience level of the user. At one end it exposes through a JavaScript frontend the whole set of functionalities available on the platform. That perfectly follows the native-web technology convergence trend exposed in section 2. At the other end, this subsystem interacts with the Mobility Manager in order to flexibly tune the QoS parameters satisfying applications’ and users’ requirements.

The “off-device” interface subsystem presents an externally accessible interface which supports external services for device management and configuration (i.e., read status and features, switch access network, read QoS statistics etc.). Off-device design and implementation will follow the OMA DM specification, and will be based on the RESTful technology.

The Mobility Manager interface is aimed to connect the FI-WARE S3C (Service Capability Connectivity & Control) system, and exploits its exposed APIs toward the EPC (Evolved Packet Core). Mobility Manager interactions with S3C are aimed to configure network policies, routing rules, and flexibly allocating bandwidth on the access network, to fulfil applications' QoS requirements.

A. *Technical requirements*

Even though the set of functionalities exposed by the different blocks (on-device, off-device and Mobility Manager) are quite different, all the described subsystems share the following technical requirements.

API discovery and linkage to application/service: Most of the identified functionalities are already provided at native level, but some of them (i.e., QoS/QoE) are not standard, and require an external native implementation. Moreover also standard functionalities could not be supported by a given device, entailing a strong necessity for an integrated API discovery service. That enables applications (local or remote) and external services to discover all APIs they depend on, for a given target platform. If APIs are successfully discovered the application/service should be able to bind and finally use the required functionalities.

API call delivery at native level: API invocations generated at API level must be conveyed down at native level. The same holds for API return values (i.e., JavaScript callbacks) that must be replied back from the native level to the API level.

In order to address these requirements, CDI selects three base platforms (one for each product release) which can fulfil them. This paper focuses on the implementation done for the CDI release which adopts the Webinos framework as a platform enabling to embed the aforementioned APIs inside the device, and subsequently publish find/bind, compose, and use them. Webinos is a web based application platform that runs over multiple devices (phone, PC, tablet, TV, vehicle etc.). Webinos runs an overlay network of web applications that use APIs offered by local and remote devices. It is based on the concept of Personal Zone (PZ), as a set of Webinos enabled devices owned by a Webinos user, and federated under a Personal Zone Hub (PZH) hosted in the cloud, or on top of one selected device in the PZ. Local and remote (intra/inter PZ) service invocation is supported by JSON-RPC 2.0 calls, and routed by Personal Zone Proxies (PZP) running on each device.

B. *CDI APIs*

Web applications and remote cloud based services, should be able to access native functionalities offered by connected devices. To this aim CDI implements a subset of the W3C API specifications , and also extends them with new and innovative functionalities. API functionalities are organized into groups, and each group is further subdivided into functional blocks:

- Device Sensors: All connected devices targeted in FI-WARE contain a range of sensors which can provide useful functionality for application developers. CDI will support the following four sensor types: Camera, Microphone, Geo-Location, Device Orientation & Accelerometer.
- Quality of Experience (QoE): CDI provides to application developers useful QoE control functionalities. The QoE control functionalities are embedded inside the "on-device" subsystem, which is able to combine explicit QoE feedback coming from the user, with the QoS level provided by the network [4]. Both inputs are matched against a target QoE level to be achieved, and network resources are consequently requested in order to reach the target level. This API has not been identified by the W3C, but it is the outcome of the original work of the CDI consortium.
- Quality of Service (QoS): CDI provides a QoS API aimed to provide access network selection functionalities, allowing devices to select network interfaces according to the connectivity requirements of applications or by network operator policies. This access network selection is transparent to the applications, that just need to express their current QoS needs. Again this API has not been identified by the W3C, but it is the outcome of the original work of the CDI consortium.
- User Profile: This set of interfaces enables applications to access identification and authentication procedures before of managing user's private and sensible data (photo, contacts, mails etc.).
- Device Feature: This API group enables on device applications and remote services to tailor their behaviour in function of the specific device features. Device Feature functionalities are then divided into two segments, "On Device Profile Information" (device form factor, screen size, CPU, Disk Space etc.) to be accessed by local applications and "Off Device Profile Information" (Media Services Support For Consumption, Current Connectivity etc.) to be accessed by remote applications and services
- Media Services: Set of APIs to discover media types and codecs supported by the device, for media consumption and production.
- Personal Data Services: Aimed at providing functionality to access (Read/Write) personal data on the device, opening up a wide range of possible applications. Many functional blocks are expected for this functional group: Contacts, Calendar, Gallery, File System access and Personal Data Service Discovery (to verify native functionality support).
- Phone: Allows to discover if the device is a phone, and in such case to access basic phone information (Engaged, Ringing, Dialling, Calling or Idle)and operations (make call, end call, answer call, reject call).
- Messaging: More and more devices are equipped with communication technology. It makes sense to offer that functionality to application developers where possible. This group of functionality deals with Email, SMS and MMS messaging.
- Device Connectivity: This functional group enables developers to access device connectivity capabilities such as detect connectivity technologies (WiFi, 3G, Bluetooth, NFC), switch connection between available communication interfaces, release connections, access connectivity features, enumerate access networks, connect to a network.

IV. CDI IMPLEMENTATION

The CDI development team is devoting a big effort in implementing W3C specifications, and investigating, validating and proposing the standardization of new APIs, that are expected to be valid enablers for proliferation of Future Internet services and applications. CDI's roadmap spans three years of activity, with three major delivery milestones as planned by FI-WARE. For each year, a base platform has been selected to be integrated with the CDI layers: PhoneGap, Webinos and Tizen . The first release was defined with PhoneGap as the underlying integration platform, as it provides a bidirectional channel between JavaScript and device's low level core. Such a channel is the only required mean to provide on-device functionalities, whose support is the one expected for the first and second FI-WARE releases. APIs integrated on PhoneGap were subsets of Device-Sensors and Personal-Data-Services, with essential (stub) support for Mobility Manager and QoE-Engine APIs. For such reason the first version is only available for the internal use of the team, and represents the starting point to extend, enhance and provide the expected functionalities for the second delivery of FI-WARE (see project site for up to date information on the project web site). Such delivery will have a working subset of the described on-device and Mobility Manager APIs, including QoE control, QoS control with basic Mobility Manager-EPC interactions, and subsets of the "Device Sensors" and "Personal Data Services" functional groups.

All these functionalities will be supported by the subsystem in Figure 2, hosted on top of the connected device, and perfectly integrated within the Webinos on device framework.

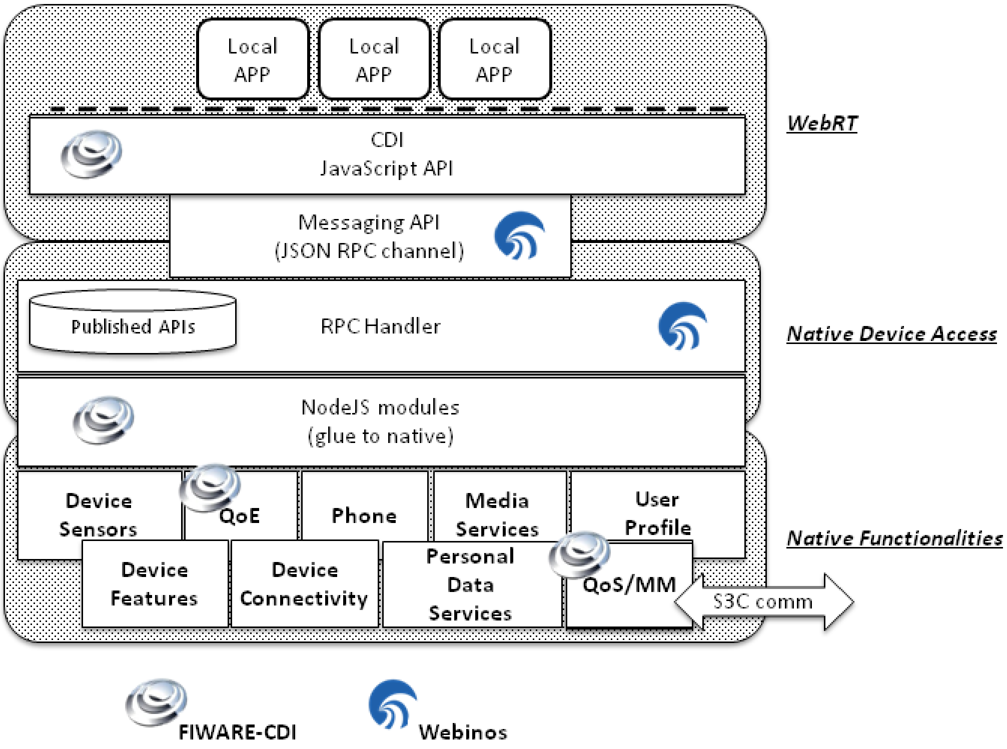


Figure 2 - CDI-Webinos integration

A. *Webinos integration*

On device web applications hosted in the browser or in the Webinos WebRT, are able to discover APIs registered by a RPC (Remote Procedure Call) handler. For the sake of brevity, in the rest of discussion we are going to refer WebRT and web browsers simply as WebRTs, with the aim to refer environments able to host and execute applications built on top of web languages. Available APIs are registered and discoverable through an URL that uniquely identifies the group of functionalities provided (i.e., http://fiware.cdi.dev-features.org/). Accordingly to the Webinos approach, for developers' use such URLs are all listed in a publicly available repository on the web. Once APIs have been discovered, Webinos offers functions to bind them to applications and finally invoke the provided methods. API Discovery/Bind/Usage invocations from applications are all packed into JSON RPC 2.0 messages, and sent by the JavaScript CDI API to a RPC handler waiting at the other end of a communication pipe. This is the first step to convey API calls out of the WebRT. The communication pipe is also engaged by JavaScript callbacks, when they are sent back to applications in response to API invocations. The whole system on the other side of the pipe is hosted by NodeJS, a highly modular JavaScript based environment, that allows multithreading, and enables access to native device libs. NodeJS is thought to provide server side functionalities and is the core platform for Webinos operations. The RPC handler is hosted as a NodeJS module, aimed to deliver RPCs to a backend CDI layer (again composed of NodeJS modules), whose purpose is to relay API calls at native level, using primitive NodeJS glue mechanisms. In the case of native Java code to be reached (i.e., in Android platforms) NodeJS does not support the mapping, but Webinos comes in help by providing a Java-bridge facility. At the lowest level in the picture we have the native platform along with all described groups of functionalities, that in some cases are not implemented by the OS and need to be separately ported into the system (i.e., QoE and QoS/Mobility Manager functionalities). For what concerns the Mobility Manager interface toward the S3C, this is implemented as a RESTful web service, and is completely independent by Webinos and its framework.

Business benefits

The CDI represents a valid solution to the fragmentation problem. Thus it is a breakthrough milestone in the Future Internet convergence path. The main effect of such convergence is measurable in the context of interoperability among the different actors of the Internet business: users, device manufacturers, network operators, content providers, application and service developers. The business impact is obvious for all the above mentioned stakeholders. Users of such devices can enjoy any available content or service through the accessible networks. Device manufacturers can use standard APIs to open their platforms to a variety of contents and services guaranteeing a continuity in the user experience and expectations. The portability of contents and services enables the "create once, run anywhere" paradigm that allows developers of applications and services as well as producers of contents to reach larger user communities. Another important aspect for the CDI acceptance is an explicit support to remote management and control of device configuration, allowing the network operator to play a key role in the cooperation between users, devices, services and contents with the Internet infrastructure. For all these reasons we believe that the CDI is a technological enabler of the Future Internet business.

V. Conclusions and Future Works

The work described aims to be a solution to the problem of platforms fragmentation that affects application portability, interoperability among heterogeneous systems and impacts heavily on remote devices configuration. Adoption of the CDI concept dramatically pushes connected devices to become flexible, ready to use, multi-purpose and vendor independent open boxes. That feeds related ecosystems with a breeze of innovation, breaking barriers for new services, applications and then revenues. The CDI team is currently reaching its 2nd milestone by integrating an important set of the identified functionalities on top of Webinos. By the time this milestone is reached, the architecture design will be improved to support extended use cases, and interfaces with other FIWARE GEs. Such interactions and dependencies are planned to be supported by the 3rd release of the project, together with complete and enhanced on-Device, off-Device and Mobility Manager functionalities, entailing also more complex interactions with the Evolved Packet Core through the FI-WARE S3C interface.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement n° 285248 also referred as FI-WARE. It is worth to mention the valuable contribution of Chris Woods, Frank Schulze, Prof. Francesco Delli Priscoli, Prof. Stefano Battilotti, Prof. Salvatore Monaco, Prof. Roberto Cusani, Dr. Antonio Pietrabissa and Dr. Vincenzo Suraci.

References

[1] Christy Pettey, "Top 10 Strategic Technology Trends for 2013", Gartner Symposium/ITxpo, Orlando, Fla., October 23, 2012

[2] Castrucci M., "Key Concepts for the Future Internet Architecture". In: Paul Cunningham and Miriam Cunningham (Eds). Future Network and MobileSummit 2011 Conference Proceedings. Warsaw, Poland, 15 - 17 June 2011, IIMC International Information Management Corporation, ISBN: 9781905824236

[3] Castrucci M. "A Cognitive Future Internet Architecture". In: Domingue, J.; Galis, A.; Gavras, A.; Zahariadis, T.; Lambert, D.; Cleary, F.; Daras, P.; Krco, S.; Müller, H.; Li, M.-S.; Schaffers, H.; Lotz, V.; Alvarez, F.; Stiller, B.; Karnouskos, S.; Avessta, S.; Nilsson, M. (Eds.). The Future Internet. p. 91-102, Springer, ISBN: 9783642208973, doi: 10.1007/978-3-642-20898-0_7

[4] Iannone M. (2012). Modelling Quality of Experience in Future Internet Networks. In: Proceedings of Future Network & Mobile Summit. p. 1-9, ISBN: 9781467303200, Berlin, 4 - 6 July 2012

[5] A. Pietrabissa, "An Alternative LP Formulation of the Admission Control Problem in Multi-Class Networks", *IEEE Transaction on Automatic Control*, Vol. 53, N. 3, pp. 839-845, April 2008, DOI: 10.1109/TAC.2008.919516.

[6] A. Pietrabissa, "Admission Control in UMTS Networks based on Approximate Dynamic Programming", *European Journal of Control*, Vol. 14, N. 1, pp. 62-75 , January 2008, DOI:10.3166/ejc.14.62-75.

[7] C. Bruni, F. Delli Priscoli, G. Koch, A. Pietrabissa, L. Pimpinella, "Network decomposition and multi-path routing optimal control", *Transactions on Emerging Telecommunications Technologies (John Wiley & Sons, Inc., USA)*, Vol. 24, Issue: 2, March 2013, pp. 154-165, doi: 10.1002/ett.2536.

[8] G. Oddi A. Pietrabissa,, F. Delli Priscoli, V. Suraci, "A decentralized load balancing algorithm for heterogeneous wireless access networks", World Telecommunication Congress (WTC), Berlin, June 2014.

[9] Delli Priscoli, F., & Isidori, A. (2005). A control-engineering approach to integrated congestion control and scheduling in wireless local area networks. *Control Engineering Practice*, *13*(5), 541-558.

[10] Cusani, R., Priscoli, F. D., Ferrari, G., & Torregiani, M. (2002). A novel MAC and Scheduling strategy to guarantee QoS for the new-generation WIND-FLEX wireless LAN. *IEEE wireless communications*, *9*(3), 46.

[11] Migliardi M., "Improving energy efficiency in distributed intrusion detection systems" (2013) Journal of High Speed Networks, 19 (3), pp. 251-264.

[12] Merlo A., "Secure cooperative access control on grid", (2013) Future Generation Computer Systems, 29 (2), pp. 497-508.

[13] Curti M., "Towards energy-aware intrusion detection systems on mobile devices" in Proc. of the 2013 International Conference on High Performance Computing and Simulation, HPCS 2013, art. no. 6641428, pp. 289-296.

[14] Mignanti S. (2007). Context-aware Semantic Service Discovery. In: Proocedings of Mobile and Wireless Communications Summit. Budapest, July 2007, p. 1-5, ISBN: 963-8111-66-6, doi: 10.1109/ISTMWC.2007.4299110

[15] Di Giorgio A. (2008). A Model Based RL Admission Control Algorithm for Next Generation Networks. In: AlBegain K; Cuevas A. Proceedings - The 2nd International Conference on Next Generation Mobile Applications, Services, and Technologies, NGMAST 2008. Cardiff, WALES, SEP 16-19, 2008, ISBN: 978-0-7695-3333-9, doi: 10.1109/NGMAST.2008.19

[16] Javaudin JP (2008). Towards convergent Gigabit Home Networks. In: IEEE PIMRC 2008 - Gigabit Home Access Special Session. Cannes, France, Sept 15-18, p. 1-5, ISBN: 978-1-4244-2643-0

[17] Fiaschetti A. (2012). The SHIELD Framework: how to control Security, Privacy and Dependability in Complex Systems. In: IEEE Workshop on Complexity in Engineering. Aachen, June 11-13, 2012, doi: 10.1109/CompEng.2012.6242962

[18] Macone D. (2013). A dynamic load balancing algorithm for Quality of Service and mobility management in next generation home networks. TELECOMMUNICATION SYSTEMS, vol. 53, p. 265-283, ISSN: 1018-4864, doi: 10.1007/s11235-013-9697-y

[19] Palo A. (2013). A common open interface to programmatically control and supervise open networks in the future internet. In: 2013 Future

Network and Mobile Summit, FutureNetworkSummit 2013. 6633572, ISBN: 978-190582437-3, Lisbon, 3 July 2013 through 5 July 2013

[20] Oddi G. (2013). A resource allocation algorithm of multi-cloud resources based on Markov Decision Process. In: Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom. vol. 1, p. 130-135, San Diego (CA), 17th-21th July 2005, doi: 10.1109/CloudCom.2013.24

[21] Marucci A. (2013). Energy-aware control of home networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 7972, p. 299-311, Berlin:© Springer-Verlag Berlin Heidelberg, ISBN: 978-364239642-7, Ho Chi Minh City; Viet Nam, 24 June 2013 through 27 June 2013, doi: 10.1007/978-3-642-39643-4_23

[22] Delli Priscoli F., "Dynamic Class of Service Mapping for Quality of Experience Control in Future Networks", World Telecommunication Congress (WTC), Berlin, June 2014.

Endoscopic Procedures Control Using Speech Recognition

Simão Afonso, Isabel Laranjo, Joel Braga, Victor Alves, José Neves

***Abstract* —** In this paper it is presented a solution for replacing the current endoscopic exams control mechanisms. This kind of exams require the gastroenterologist to perform a complex procedure, using both hands simultaneously, to manipulate the endoscope's buttons and using the foot to press a pedal in order to perform simple tasks such as capturing frames. The last procedure cannot be accomplished in real-time because the gastroenterologist needs to press an additional programmable button on the endoscope to freeze the image and then press the pedal to capture and save the frame. The presented solution replaces the pedal with a hands-free voice control module and it is capable of running on the background continuously without human physical intervention. This system was designed to be used seamlessly with the MyEndoscopy system that is being tested in some healthcare institution and uses the PocketSphinx libraries to perform real-time recognition of a small vocabulary in two different languages, namely English and Portuguese.

***Keywords*—**Automatic Speech Recognition, Hidden Markov Models, PocketSphinx, SphinxTrain, Endoscopic Procedures

I. Introduction

NOWADAYS it is accepted by most healthcare professionals that information technologies and informatics are crucial tools to enable a better healthcare practice. The Pew Health Professions Commission (PHPC) recommended that all healthcare professionals should be able to use information technologies [1]. The technological evolution has led to an enormous increase in the production of objective diagnostic tests and a decrease on the reliance of more subjective problem solving methods, which should increase the quality of the service provided, and can even be seen as a consequence of the increased accountability of healthcare institutions in relation to the legislation [2].

EsophagoGastroDuodenoscopy (EGD) and Colonoscopy occupy relevant positions amongst diagnostic tests, since they combine low cost and good medical results. The current endoscopic exams require the gastroenterologist to perform a complex procedure using both hands simultaneously to manipulate the endoscope's buttons and using the foot to press the pedal in order to perform such simple tasks as capturing frames. The last procedure cannot be accomplished in real-time because the gastroenterologist needs to press an additional programmable button on the endoscope to freeze the image and then press the pedal to capture and save the frame [3]. This approach to the problem is not optimal and raises several new issues, such as limiting the movements of everyone involved and requiring the gastroenterologist to perform a complex procedure, distracting him/her from the task at hand. A new hands-free interface that allows for a richer control scheme would solve some of the existing snags.

A novel approach to this problem consists of adding a voice recognition module to the system, providing a hands-free control. This module, called *MIVcontrol*, will be integrated into the device called *MIVbox* (more details are given in section 3).

The main goal of the *MIVcontrol* module is to create a simple speech recognition system for recognizing a very small vocabulary of simple pre-determined commands. The recognized commands are used to control the *MIVacquisition*, creating a hands-free control system that should be able to replace the current solution. This system can perform frame capturing in real-time, without the need to use any extra buttons.

The system should be speaker-independent and have a very low error rate, even on noisy environments, and it should be able to capture audio from a microphone continuously, so that it can run in the background without human intervention. This will require automatic word segmentation, to make recognition possible.

The rest of the paper is organized as follows: in section 2 it is presented a review of related work in the area of speech recognition, from its theoretical foundations to practical systems already being used. In section 3 is outlined the overall system architecture and how it integrates with the *MyEndoscopy* system. In section 4 is presented specific details about the implementation of the solution, whereas in section 5 the methodology used in the study is exhibited. Finally, the results and their assessment are presented in section 6 and 7, followed by conclusions in section 8.

II. Related Work

Automatic Speech Recognition (ASR) is a process by which a computer processes human speech, creating a textual representation of the spoken words. This process has two main areas of study, i.e. discrete speech and continuous speech. Discrete speech is useful for the creation of voice command interfaces, while continuous speech, also known as dictation, mimics the way two humans communicate. Though the ultimate objective of having a system capable of

This work is funded by ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT - *Fundação para a Ciência e a Tecnologia* (Portuguese Foundation for Science and Technology) within project PEst-OE/EEI/UI0752/2014.

S. Afonso, I. Laranjo, J. Braga and J. Neves are in the Computer Science and Technology Center (CCTC), University of Minho, Braga, Portugal (simaopoafonso@gmail.com, isabel@di.uminho.pt, jneves@di.uminho.pt, joeltelesbraga@gmail.com)

V. Alves is in the Computer Science and Technology Center (CCTC), University of Minho, Braga, Portugal (corresponding author to provide e-mail: valves@di.uminho.pt).

recognizing everything anyone can say in multiple languages has yet to be achieved, research has been focused on smaller-scale approaches [4].

A. Theoretical Foundations

Aymen et al. [5] presented the theoretical foundation of Hidden Markov Models (HMM) that underpin most modern implementations of automatic speech recognition. The authors present the distinction between speech recognition, which aims to recognize almost anyone's speech, and voice recognition, which creates systems trained to particular users. The model is constructed based on a large *corpus* of recorded speech, annotated with the respective transcription. The HMM requires three different sub-models:

1) The acoustic model consists of different features for each utterance the system recognizes;
2) The lexical model tries to identify sounds considering the context;
3) The language model identifies the higher-level characteristics of speech, such as words and sentences.

The HMM searches the model for similar patterns that fit into the given audio input, producing probable matches. The HMM's advantages over previous learning algorithms consists of easy implementation on a computer and automated training without human intervention. This stems from the fact that it is assumed that in short-time ranges the process is stationary, vastly reducing the computational effort [5].

B. Implementations

There are several HMM implementations, but the most advanced are the HTK Toolkit [6] and the CMU Sphinx system [7].

Hidden Markov Model Toolkit (HTK) is a set of libraries used for research in automatic speech recognition, implemented using HMM. The HTK codebase is owned by Microsoft, but managed by the Cambridge University Engineering Department. Since HTK has been largely abandoned, since the last release (v3.4.1) was made in 2009, the CMU Sphinx system is getting more attention from the speech recognition community [6].

The original SPHINX was the first accurate Large Vocabulary Continuous Speech Recognition (LVCSR) system, using HMM as its underlying technology, that managed to be speaker independent [7]. The next version, SPHINX-II, was an improved version that was both faster and more accurate, created by most of the same authors, using HMM as its underlying technology. It was developed, from the beginning, as an open source project, creating a community around it [8]. The next version, SPHINX-III, is an offline version of the previous systems, with a different internal representation to allow for greater accuracy. The signals go through a much larger amount of pre-processing before they even reach the recognizer [7]. Current hardware is capable of running the recognizer for SPHINX-III in almost real-time, but it is not suitable to processing in such conditions. SPHINX-4 is a complete rewrite to create a more modular and flexible system that can accept multiple data sources elegantly. It is a joint venture with Mitsubishi Electric Research Laboratories and Sun Microsystems, using the Java programming language. As with the third version, its intended use is offline processing, not real-time applications [9]. *Vertanen* [10] tested both the HTK and the Sphinx systems with the Wall Street Journal (WSJ) corpus and found no significant differences in error rate and speed. This conclusion is corroborated by other researchers [11].

Huggins-Daines et al. [12] optimized CMU Sphinx II for embedded systems, primarily those with ARM architecture. To balance the loss of precision required in other optimizations, the CMU Sphinx III Gaussian mixture model was back-ported. They managed to have a 1000-word vocabulary running at 0.87 *times real-time* on a 206 MHz embedded device, with an error rate of 13.95% [12]. "*Times real-time*" is a notation that indicates the amount of time required to process live data. In this case, the system can process 1 second of data in 0.87 seconds, which makes it suitable to real-time recognizing. This work has lead to the creation of the PocketSphinx project, an open source initiative to continue this work. This project is in active development, and it has bindings for C and Python [13].

C. Practical systems

Vijay [14] studied the problem of phonetic decomposition in lesser-studied languages, like Native American and Roma language variants, using the PocketSphinx system. While the system does not implement the complex rules of these languages, it is possible to leverage the existing system to recognize unknown languages, using a relatively simple lookup table that maps sounds to phones [15]. *Varela et al.* [15] adapted the system to the Mexican Spanish language. The authors created a language and an acoustic model, based on an auto-attendant telephonic system, and achieved an error rate of 6.32% [15]. The same process was followed for other languages, like Mandarin [16], Arabic [17], Swedish [18]. These examples show that the PocketSphinx system is flexible enough so that it is relatively easy for people with phonetics training to extend it to other languages.

Harvey et al. [4] researched how ASR systems could be integrated with their project aimed at developing a device to help the elderly, both inside and outside the home. The authors identified the following challenges associated with ASR systems used for voice command interfaces [4]:

1) Important differences between users;
2) Similarity between certain sounds;
3) Short words provide less data for the system to analyze, which may lead to increased error rates;
4) Different recognition languages lead to variable error rates using the same system.

Specifically to their project, the authors found that medical conditions, which frequently affect the elderly, create different speech patterns and their tolerance to errors is quite low. With that in mind, the authors leveraged the Sphinx library for its maturity and features. Focusing on the creation of models and general optimization tasks, the authors managed to create a multilingual system that has a 2-second processing time on embedded systems, with error rates above 70% [4].

Kirchhoff et al. [19] suggested other methods to improve the ASR systems' performance. One proposal consists of replacing the current feature-extraction algorithms with others specially designed to discriminate certain sound classes depending on the intended use, or through the use of noise reduction algorithms, which can improve the data

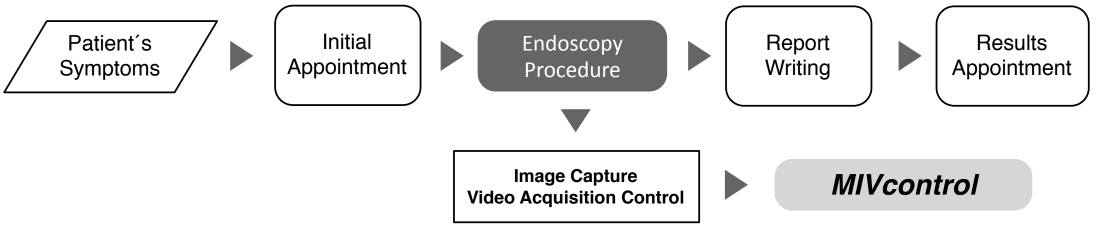


Fig. 1 Workflow for a gastroenterology medical appointment

analysis and increase the models' accuracy, using the same data collection routines. A different approach is collecting more out-of-band information to increase the amount of data available to the system. This might include different processing front-ends for feature extraction (although this may be of limited use) or even non-acoustic data, such as visual information [19].

III. Architecture

As referred before, the gastroenterologist needs to use a pedal to capture and save the frame, and with the proposed solution the pedal is replaced with a hands-free voice control module, called *MIVcontrol*. This module was developed to tackle the problems that healthcare professionals face when performing an endoscopic procedure. This module is part of the *MIVbox* device, which is integrated in the *MyEndoscopy* system.

MyEndoscopy is the name of the global system developed by *Laranjo et al.* [21], which groups several *MIVboxes*, which are scattered by various healthcare institutions. The main goal of *MyEndoscopy* is to link different entities and standardize the patient's clinical process management, to promote the sharing of information between different entities [21].

The *MIVbox* device has a web-based distributed architecture and it is capable of acquisition, processing, archiving and diffusion of endoscopic procedure results [21]. The main goal of the *MIVcontrol* module is to replace the pedal, currently used by gastroenterologists to capture interesting frames, by voice commands that interact directly with the *MIVacquisition* module. The MIVacquisition module receives the video directly from the endoscopic tower and provides it to all the MIVbox modules [20].

In **Fig. 1** is presented a simple workflow describing the moments that occur in a gastroenterology medical appointment, in the healthcare institution, that results in an Endoscopy Procedure. The *MIVcontrol* module can be seamlessly integrated into the current workflow by allowing the gastroenterologist to control the *MIVacquisition* module by using voice commands during the endoscopic procedure.

In **Fig. 2** is presented the overall system architecture, as a component of the *MIVbox* device. The *MIVcontrol* module uses the live audio streaming from a microphone to recognize commands and send them to the *MIVacquisition* module.

The process that leads to the creation of a model is presented on **Fig. 3**. It uses a corpus of pre-labeled audio data to create a speech model that can be used in the *MIVcontrol* module. This speech model is a combination of acoustic and language models.

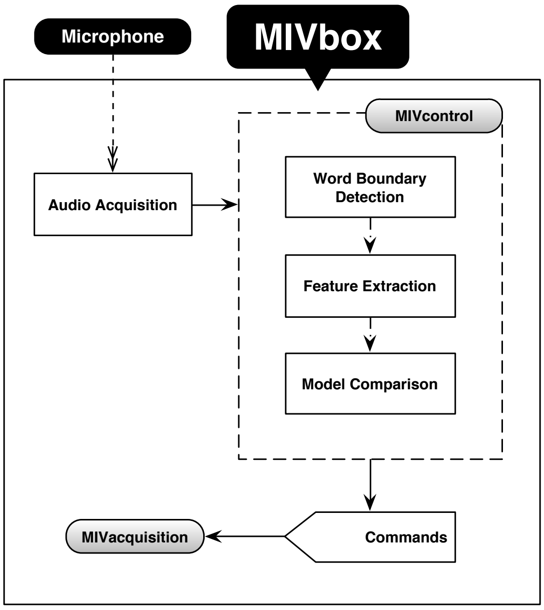


Fig. 2 *MIVcontrol* global architecture

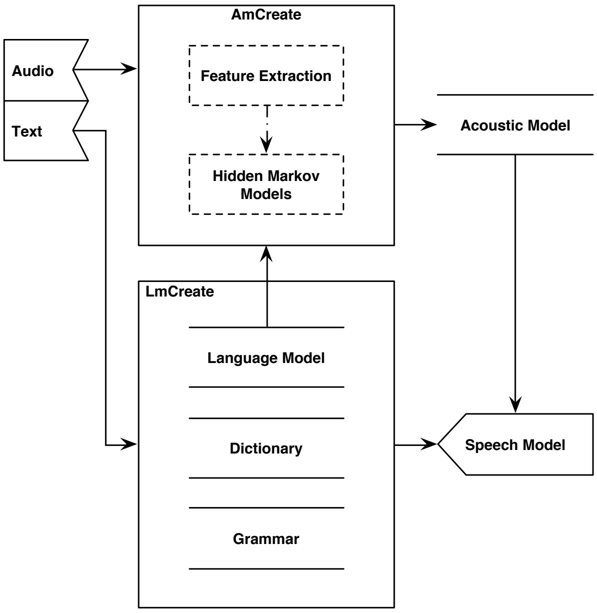


Fig. 3 *MIVcontrol* model training procedure

The process is split in two main sub-processes: creation of the textual model, named *lmCreate*, and the creation of the acoustic model, named *amCreate*. Since the acoustic model requires parts of the textual model, it must be generated last. The *lmCreate* process creates the textual model based on the textual data in the corpus, while the *amCreate* process analyzes the pre-recorded audio data using the same feature extraction steps used in the *MIVcontrol* module, and uses HMM to learn how to classify the commands contained in the language model.

IV. Implementation

The creation of the speech model used in the *MIVcontrol* module, from higher to lower level comprises three different phases, namely language model, dictionary and acoustic model.

A. Language Model

The language model is a high-level description of all valid phrases (i.e. combination of words) in a certain language. Statistical language models try to predict all the valid utterances in a language, by combining all the recognized words into every possible combination [22]. Context-Free Grammars are restricted forms of a language model, that restrict the recognized phrases to a predetermined set, and discard those that do not fit that model [23].

The decision to adopt a certain language model depends mostly on its intended application. While statistical language models are useful for open-ended applications, like dictation and general-purpose recognition, context-free grammars are suitable for specific applications, like command-and-control systems.

SphinxBase requires the grammar to be defined in Java Speech Grammar Format (JSGF), which is a platform-independent standard format to define context-free grammars, using a textual representation so that it can be human-readable [24]. The statistical language model is automatically created based on the command list.

B. Dictionary

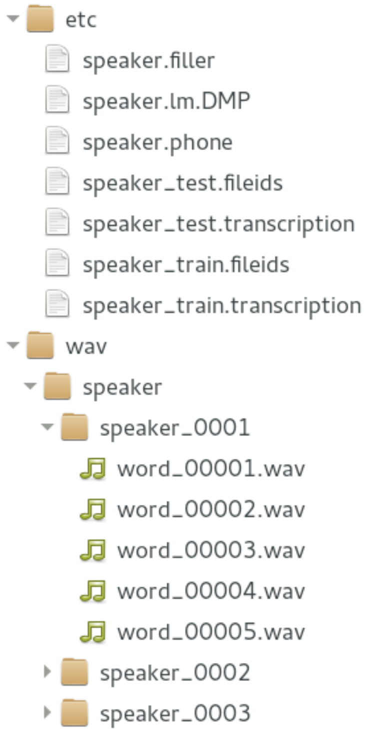
The dictionary is a map between each command and the phonemes it contains. A phoneme is defined as the basic unit of phonology, which can be combined to form words. Its internal representation consists of using the ARPAbet to represent phonemes as ASCII characters. The ARPAbet does not allow representing the entire International Phonetic Alphabet (IPA), but it is sufficient for small vocabularies, such as the one required by this application [25].

Since the list of required commands is small, all the dictionaries used were created manually.

C. Acoustic Model

The acoustic model is trained using *SphinxTrain* and maps audio features to the phonemes they represent, for those included in the dictionary. The training performed by SphinxTrain requires previous knowledge of the dictionary and a transcription for each utterance, in order to map each utterance to its corresponding phonetic information. It also requires the data to be in a particular audio format. In order to minimize clerical errors and cut the time need to analyze the data to a minimum, all the technical considerations and index building were abstracted away in a script referred to as *amCreate*.

SphinxTrain requires the folder tree presented on **Fig. 4**, where "**model**" denotes the model name.



```
etc
  speaker.filler
  speaker.lm.DMP
  speaker.phone
  speaker_test.fileids
  speaker_test.transcription
  speaker_train.fileids
  speaker_train.transcription
wav
  speaker
    speaker_0001
      word_00001.wav
      word_00002.wav
      word_00003.wav
      word_00004.wav
      word_00005.wav
    speaker_0002
    speaker_0003
```

Fig. 4 Folder tree required by *SphinxTrain.*

The folder directory has two top folders, namely the *etc* and the *wav*.

The *etc* folder contains all the metadata and configuration parameters needed to train the acoustic model, as well as the dictionary. It contains both a list of all the phonemes used in the model and a list of filler phonemes, such as silences, that should be ignored. It also has a list of all the files to be used during both training and testing phases, as well as a mapping between each audio file and its corresponding transcription. This mapping corresponds to the labeled data to be the input to the HMM.

The *wav* folder simply contains all the collected data, as audio files, organized in subfolders by speaker identification, with a subfolder for each set of uttered commands.

The system processes continuous audio in real-time, splits it in commands and produces a line of text for each recognized command. If the spoken command is not recognized, an empty line is produced. The *MIVcontrol* module runs on the *MIVbox*.

The audio picked up by the microphone is stored in a memory buffer. The first pre-processing stage involves splitting the incoming audio into different utterances, or sets of words, by tracking silent periods between them. To account for noise present during recording, any audio with volume below a certain threshold is considered a silence.

Each segmented utterance then goes through a similar process. The audio is processed creating a set of features, and then the Semi-Continuous HMM finds the most likely utterance contained in its dictionary. This is the final output, corresponding to a command given to the system.

If there is Internet access and the data to be recognized is not sensitive, it is possible to use an online speech recognition service, such as the Google Speech API [26], as a fallback mechanism.

V. Methodologies

The parameters that have a bigger impact on the model's accuracy are the number of tied states used in the HMM and the number of Gaussian mixture distributions, so testing will focus on this parameters. Before testing begins, the data is split randomly between training and testing stacks, with the testing stack receiving 10% of the data. That same data is tested varying both the number of tied states in the HMM and the number of Gaussian distributions. The accuracy of the model is represented as a Word Error Rate (WER), which combines both false positives and false negatives into a single metric. This is done because this module acts as *middleware*, being used by other modules on a global system. The context where the commands are spoken can then be considered, which is not evaluated here. Furthermore, this methodology is consistent with the literature on the subject.

The results were obtained on a computer with 2 GB of RAM and an Intel Celeron CPU, with two cores and a clock speed of 1.10GHz. The operative system used was Fedora 19, 32 bits version. The compiler used was gcc v4.8.2, using PocketSphinx v0.8 and SphinxBase v0.8, both from the official repositories. The training used CMUcltk v0.7, compiled from source, and SphinxTrain v1.0.8, from the official repositories. The training data was collected with the built-in laptop microphone, in both noisy and quiet conditions, to better correspond to the concrete use-case.

VI. Results

The audio corpus in which the system was tested contained two languages, Portuguese and English, with a total of 1405 recordings, totaling 25 minutes of speech, recorded by 5 female and 7 male speakers. To test this model, *SphinxTrain* tried to predict the contents of the testing data using the model created with the training data. The main parameters that can be tweaked are the number of tied states in the HMM and the number of Gaussian mixtures distributions.

The effect of the number of tied states in the HMM is shown on **Table 1**.

Table 1 Effect of the number of tied states on the WER

Number of Tied States	Errors	WER (%)
5	112	33.6
10	85	28.7
25	109	32.2
50	100	35.6
100	86	26.8
150	108	33.3

The effect of the number of Gaussian mixtures distributions on the error rate is shown on **Table 2**.

Since the trained model is small, the differences in processing time are negligible. That defined the optimal conditions for training 8 Gaussian mixture distributions with 100 tied states in the HMM. With this configuration, the system classified the Portuguese model with 11.22 % WER and the English model with 4.55 % WER.

Table 2 Effect of the number of Gaussians on the WER

Number of Gaussians	Errors	WER (%)
1	152	17.1
2	172	17.0
4	134	14.6
8	142	14.1

VII. Discussion

As a proof-of-concept, this system managed to create a voice recognizer for a very small vocabulary to be used as a command and control system, leveraging the capabilities of the CMU Sphinx project. It was created as an alternative to cloud-based solutions, such as Google Speech API. In a medical environment, cloud-based solutions pose certain challenges that might degrade their performance, such as increased communications security, a need to keep recurring costs on non-medical equipment to a minimum, and also privacy and legal reasons on systems that deal with sensitive data. Having a system that can be installed inside the healthcare institutions' network without external dependencies is a plus for the reasons presented above.

PocketSphinx was based on work done for SPHINX II, which was not designed as a real-time recognizer. With all the optimizations it has received, it is possible to use it in real-time with acceptable performance, even in underpowered computers.

VIII. Conclusion and Future Work

In summary, this paper presents an automatic speech recognition system designed specifically to solve a problem that affects gastroenterologists. The system is capable of running on the background continuously without human physical intervention, and so it is capable of replacing the pedal and buttons commonly used in current endoscopic systems. It was designed to be used seamlessly with the MyEndoscopy system that is being tested in some healthcare institutions.

The next step will involve improving the integration with the *MyEndoscopy* system, including a more robust testing phase, which is facilitated by the fact that PocketSphinx is a cross-platform library.

To increase the usefulness of the system, it is important to collect more data, particularly with different voice features. It is also possible to apply newer training algorithms to the same data, and test how they affect the models created. The *SphinxTrain* suite was created using CMU Sphinx recognizers, but there are more recent projects that are able to produce models compatible with PocketSphinx-based recognizers. Those newer systems may generate better models with the same data.

Reference

[1] E. H. O'Neil, "Recreating Health Professional Practice for a New Century," San Francisco, CA, 1998.

[2] N. Summerton, “Positive and negative factors in defensive medicine: a questionnaire study of general practitioners.,” *BMJ*, vol. 310, no. 6971, pp. 27–29, Jan. 1995.

[3] J. M. Canard, J.-C. Létard, L. Palazzo, I. Penman, and A. M. Lennon, *Gastrointestinal Endoscopy in Practice*, 1st ed. Churchill Livingstone, 2011, p. 492.

[4] A. P. Harvey, R. J. McCrindle, K. Lundqvist, and P. Parslow, “Automatic speech recognition for assistive technology devices,” in *Proc. 8th Intl Conf. Disability, Virtual Reality & Associated Technologies*, Valparaíso, 2010, pp. 273–282.

[5] M. Aymen, A. Abdelaziz, S. Halim, and H. Maaref, “Hidden Markov Models for automatic speech recognition,” in *2011 International Conference on Communications, Computing and Control Applications (CCCA)*, 2011, pp. 1–6.

[6] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, “HTK FAQ.” [Online]. Available: http://htk.eng.cam.ac.uk/docs/faq.shtml. [Accessed: 03-Feb-2014].

[7] K.-F. Lee, H.-W. Hon, and R. Reddy, “An overview of the SPHINX speech recognition system,” *IEEE Trans. Acoust.*, vol. 38, no. 1, pp. 35–45, 1990.

[8] X. Huang, F. Alleva, H.-W. Hon, M.-Y. Hwang, K.-F. Lee, and R. Rosenfeld, “The SPHINX-II speech recognition system: an overview,” *Comput. Speech Lang.*, vol. 7, no. 2, pp. 137–148, Apr. 1993.

[9] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, “The CMU SPHINX-4 speech recognition system,” in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, 2003, vol. 1, pp. 2–5.

[10] K. Vertanen, “Baseline WSJ Acoustic Models for HTK and Sphinx: Training recipes and recognition experiments,” *Cavendish Lab. Univ. Cambridge*, 2006.

[11] G. Ma, W. Zhou, J. Zheng, X. You, and W. Ye, “A comparison between HTK and SPHINX on chinese mandarin,” in *IJCAI International Joint Conference on Artificial Intelligence*, 2009, pp. 394–397.

[12] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, “Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices,” *2006 IEEE Int. Conf. Acoust. Speed Signal Process. Proc.*, vol. 1, pp. I–185–I–188, 2006.

[13] D. Huggins-Daines, “PocketSphinx v0.5 API Documentation,” 2008. [Online]. Available: http://www.speech.cs.cmu.edu/sphinx/doc/doxygen/pocketsphinx/main.html. [Accessed: 20-Feb-2014].

[14] V. John, “Phonetic decomposition for Speech Recognition of Lesser-Studied Languages,” in *Proceeding of the 2009 international workshop on Intercultural collaboration - IWIC ’09*, 2009, p. 253.

[15] A. Varela, H. Cuayáhuitl, and J. A. Nolazco-Flores, “Creating a Mexican Spanish version of the CMU Sphinx-III speech recognition system,” in *Progress in Pattern Recognition, Speech and Image Analysis*, vol. 2905, A. Sanfeliu and J. Ruiz-Shulcloper, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 251–258.

[16] Y. Wang and X. Zhang, “Realization of Mandarin continuous digits speech recognition system using Sphinx,” *2010 Int. Symp. Comput. Commun. Control Autom.*, pp. 378–380, May 2010.

[17] H. Hyassat and R. Abu Zitar, “Arabic speech recognition using SPHINX engine,” *Int. J. Speech Technol.*, vol. 9, no. 3–4, pp. 133–150, Oct. 2008.

[18] G. Salvi, “Developing acoustic models for automatic speech recognition,” 1998.

[19] K. Kirchhoff, G. A. Fink, and G. Sagerer, “Combining acoustic and articulatory feature information for robust speech recognition,” *Speech Commun.*, vol. 37, no. 3–4, pp. 303–319, Jul. 2002.

[20] J. Braga, I. Laranjo, D. Assunção, C. Rolanda, L. Lopes, J. Correia-Pinto, and V. Alves, “Endoscopic Imaging Results: Web based Solution with Video Diffusion,” *Procedia Technol.*, vol. 9, pp. 1123–1131, 2013.

[21] I. Laranjo, J. Braga, D. Assunção, A. Silva, C. Rolanda, L. Lopes, J. Correia-Pinto, and V. Alves, “Web-Based Solution for Acquisition, Processing, Archiving and Diffusion of Endoscopy Studies,” in *Distributed Computing and Artificial Intelligence*, vol. 217, Springer International Publishing, 2013, pp. 317–24.

[22] P. Clarkson and R. Rosenfeld, “Statistical language modeling using the CMU-cambridge toolkit,” in *5th European Conference on Speech Communication and Technology*, 1997, pp. 2707–2710.

[23] A. Bundy and L. Wallen, “Context-Free Grammar,” in *Catalogue of Artificial Intelligence Tools*, A. Bundy and L. Wallen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1984, pp. 22–23.

[24] A. Hunt, “JSpeech Grammar Format,” 2000.

[25] R. A. Gillman, “Automatic Verification of Hypothesized Phonemic Strings in Continuous Speech,” Arlington, Virginia, 1974.

[26] B. Ballinger, C. Allauzen, A. Gruenstein, and J. Schalkwyk, “On-Demand Language Model Interpolation for Mobile Speech Input,” *Elev. Annu. Conf. Int. Speech Commun. Assoc.*, no. September, pp. 1812–1815, 2010.

Adaptive design process for responsive web development

Zsolt Nagy

***Abstract*—**Software development process is complicated and complex. Certain parts of and relations between them require modeling in order to make the process modular and easily understandable. The most known and widespread model is the waterfall model. This model is excellent in cases when we know all the system requirements at the very beginning of the development process. However it is less and less true in real life. Developers need a method, which is quite flexible and makes it possible to modify the design and system development process and architecture based on continuous customer consultations. Such a method is Agile Software Development, although it cannot be used in many cases. We combined both design process with our own experiences and created a new development model called Adaptive Design Process.

***Keywords*—**design process, responsive web, Rich Internet Applications, web development

I. Introduction

Development of web-based system gives much more tasks to a software engineer as a traditional software development; the lifecycle of the system, the development process, support and maintenance all differs from a classic software engineering. That is why traditional software development methods are not suitable for web-based systems; more precisely they need certain corrections and extensions.

Powel famously describes the complexity of web engineering and web systems in his book:

"(Web systems) involve a mixture between print publishing and software development, between marketing and computing, between internal communications and external relations, and between art and technology" [1].

II. Current Web development models

A. Waterfall model

Web-based systems differ from classic software engineering in many ways; so true it is for the design and development process. Software development process is complicated and complex. Certain parts of and relations between them require modeling in order to make the process modular and easily understandable. Several models were developed in the past thirty years such as spiral model [2], prototype model [3] or V-model [4], however the most known and widespread model is the waterfall model.

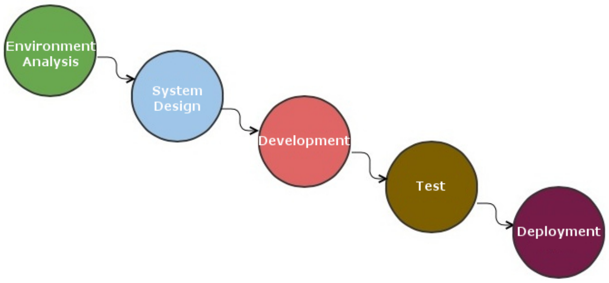


Fig. 1 Waterfall model

This model is excellent in cases when we know all the system requirements at the very beginning of the development process**.** It requires thorough preliminary survey and research work, not to mention the fact that the customer must be well prepared in the knowledge of the required functionality of the future system. Otherwise altering the design process during development is very expensive; after certain steps it is better to restart the whole process from the beginning.

However, if we own a thorough and detailed plan, both development costs and time can be well estimated. Moreover, a strict, plan-based design process is less sensitive to developer exchanges; it is easy to insert a new employee into our existing developer team. The model features that the next phase will not start until the previous phase has been completed. Thus, a developer who was delegated for a part of the process can work on a different project after finishing his current task. The system is transferred to the users only at the last stage of its life-cycle, then turns out that the software meets the initial specifications or not, the customer got what he wanted or not.

Unfortunately, in real life customers cannot define exactly what they want at the beginning of the project; requirements change and refine during development process. That is why the waterfall model is declined in many cases during business web development. The waterfall model is not prepared for this kind of changes; when a development phase has been completed, it is almost impossible to make any changes on it.

This work was supported by TÁMOP-4.2.2.A-11/1/KONV-2012-0051

Zsolt Nagy is with the Institute of Mathematics and Computer Science, College of Nyiregyhaza, Nyiregyhaza, H4400 Hungary (email: info@nagyzsolt.hu)

B. *Agile Methods*

A method is needed that is flexible, makes it possible to regularly consult with the customer and if a modification is required, it enables with the lowest cost.

Such a method is the agile software development, which is the opposite of the waterfall model in many ways. As a small module of a whole system is finished, developers immediately share it for testing and consult with the customer, who can try it, give sudden feedback, refine his needs; continuous consultation and iteration is what agile software development is based on.

This method guarantees that the user is satisfied and got what he wanted, even if he could not sufficiently formulate his demands at the beginning of the development. Adamkó (Fig. 2) describes a development process [5] similar to the agile method, which is a more modern version of Murugesan's model (Fig. 3) [6] and is still appropriate.

However, we made a revised, more detailed hybrid method to meet the needs of the demands of the responsive, intelligent web systems.

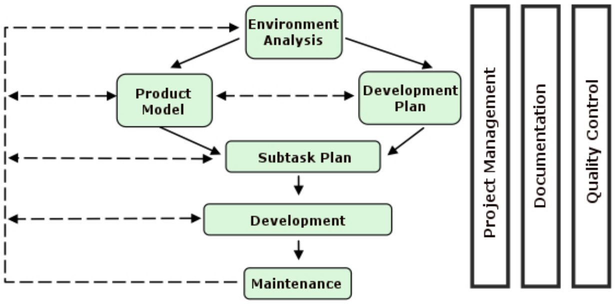


Fig. 2 Adamkó's web engineering process

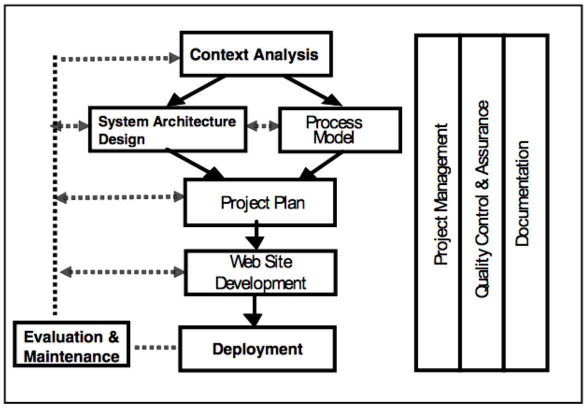


Fig. 3 Murugesan's web engineering process

.

III. Reasons for a New Model

Riding the wave of fashion of agile methodology, we could say that today's development process is clearly the agile method, however the method has several drawbacks. Continuous consultation is a great tool, but time-consuming and even if the customer does not have time for us, it pulls back the development. Due to constant revision and changing possibility, more professionals (designer, system designer, front-end, back-end developers) should be available at the same time, as opposed to the waterfall model. In addition, the expected costs and completion deadline is also difficult to estimate.

A reasonable assumption is that we can offer a more efficient method for web development by combining the benefits of the two systems. While we worked on our new method, we built in our practical experiences from the business-world and we tried to avoid the mistakes committed by us using inappropriate methodology or by just disregard any methodology in a web development process.

As the development of a user interface, design and image of a system is quite important, we built those into our model as well. Previous models elegantly passed this point or just mixed it with the system architecture design, although the so-called "look & feel" has a major importance for the customer and future users.

IV. User Interface (UI) design

A. *Traditional UI Design process*

So far a proven business practice of making a web design is based on the following process. After analyzing the needs and visiting competitors' web pages, the customer describes his ideas about the colors, the lookout, the layout and the functionality.

Based on it, a sitemap and a wireframe are created then two or three Photoshop design plans are presented to the user. The customer reviews them, chooses the right one or requires a modification, and then finally we have an accepted design plan with customer's signature. After, a HTML template is made then it will be forwarded to the software developers (Fig. 3).

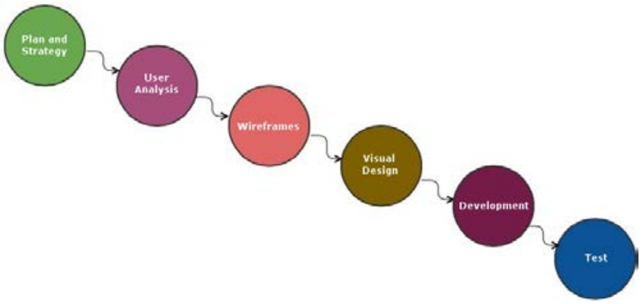


Fig. 4 Traditional design process of a user interface

The process seems familiar; yes it is exactly a classic waterfall model. In case of adaptive and responsive web systems, this process is not viable. It is impossible to create the design plan for every single device type and then making it accepted by the customer. We do not have to force this procedure onto the user or onto ourselves. Instead, a faster, iterative method should be used.

Viljami [7] and Boulton [8] describe a real-world business design process, which had an impact on creation of the following process.

B. *Responsive UI development*

In the age of responsive user interfaces (UI), we achieve the following steps after an ordinary requirement analysis and information gathering components.

1) Sketch

After finishing an information gathering process, we create a sketch, mostly using paper and pencil. Although there are very good software in the market, such a helpful one is the Zurb Responsive Sketchsheet[1]. Sketches are quick freehand drawings that are not suitable for modeling the final product; instead they play foundation roles in further design processes. Since they can be created very quickly, a sketch is a great tool to throw sudden ideas onto the canvas and show it to the user.

Literature experienced experts often encounters the concepts of sketches and wireframes that are usually confused each other. An article of UXMovement clearly describes the difference between them [9].

2) Wireframe

After finishing the sketch we can start creating a wireframe that is the architectural plan of the prospective interactive user interface. It can be considered as a visual model, however wireframes are omitted in several small and medium size projects. The justification is legitimate; sketches are fast, prototypes are slow, but interactive and informative. Wireframes are between the mentioned ones, neither fast nor informative enough for the customer.

3) Prototype

At this point, a sketch-based view is created from the combination of HTML and CSS that can be displayed in different web browsers and mostly on different type and different resolution mobile devices. Basic graphical elements appear at this stage, however detailed elaboration will be carried out only in the next step.

4) Design

Typically, this is the phase where design layouts are finished with the help of Photoshop, Fireworks or other graphical software tools. The development of graphical details is project-specific, since similar to the previous steps this is an iterative process as Fig. 5 describes it.

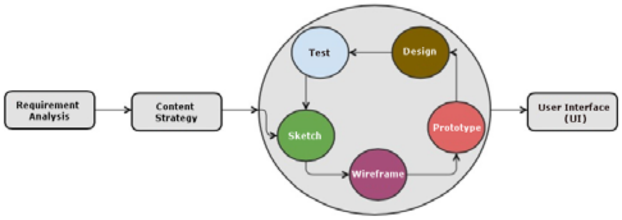


Fig. 5 Responsive UI development

It is clear that in many parts of the development process have resonance for agile software development elements. Thus, both the user interface and the whole web system development process cry for a kind of a combined solution.

V. New design process model

No matter how problematic is the waterfall model, the expected duration and the possible costs of a project can be very well estimated by using it. These two factors have high priority considering the development contract; we must fix them at the very beginning of the project.

It is also an advantage if we are able to gather as much information as possible for preparing system specification and architecture; a thorough plan greatly facilitates the future work. At the same time, the flexibility of agile methods is essential since the initial functionality always varies and expands in real-world environment.

New modules, new navigation structure, new lookout: the competitor offers new services so we should built those on our existing system; marketing experts have found out a new content layout, we should modify the existing layout; the dominant color should not be orange, but light green, we must create a whole new visual design, the customer changed his mind and he needs a geo-sensitive newsletter module and a related admin interface. All the mentioned ones are real examples, for which our development process should be prepared.

Based on the above, we offer the following process model for today's modern and adaptive web system development.

[1] http://zurb.com/playground/responsive-sketchsheets

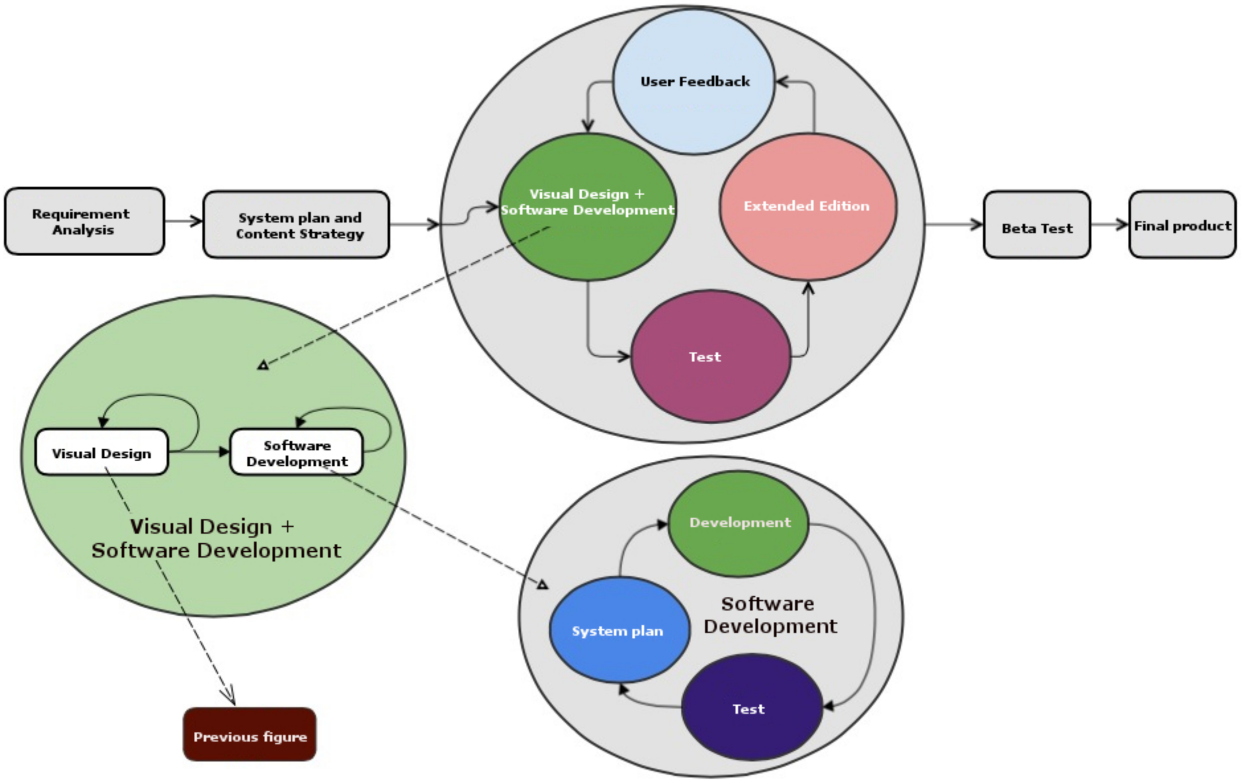


Fig. 6 Adaptive system development model

In the following paragraph we are going to describe the mentioned development phases.

A. *Requirement analysis*

Requirement analysis is the first step of web development; at this point a web engineer has to achieve the following tasks:

1. Identify the project stakeholders and assess their needs and experiences. Know the number and type of prospective users of the web system; adjust the system to these criteria and metrics.

2. Determine the services provided by the web system. This step will form the basis of menu and navigation system design and different modules of the web portal will be also determined by those information.

3. Determine what information should be displayed on the web site, how this information will be available and how often will this information change.

4. Assess the customer's needs in the area of look and feel, safety and maintenance.

5. Get to know similar websites, competitors' web systems, study their features, strengths and limitations. This step is particularly useful in an information collection process specified in paragraph 3; the customer can formulate his needs faster when he sees specific ideas and samples.

B. *System design and content strategy*

In this phase, we determine the components of the system and relationship between those components. Here we describe network and server connections (even it is a client-server, web, application or database server), system modules and their functionalities and services that these modules should implement.

This is the stage where the application of a design pattern enters; we determine the system architecture here, the user interface, data model and application logic will be developed and separated from each other.

Due to the characteristics of agile methodologies initial system functions and services may change in the future, but the design pattern, the applicable hardware and software technology and development frameworks are specified and fixed here.

Content strategy deserves special attention because of the responsiveness; not because of the content of mobile devices differs from the content of desktop computers, indeed. The golden rule of responsive development is: Mobile users want the same content as desktop users.

The idea that the view of a mobile web system is a downgraded version of a normal web portal is false. It is also a mistake if someone wants to display the same 960-pixel wide user interface content on a 320 or 480 pixel resolution smartphone.

The correct and appropriate content strategy is the “mobile-first” trend that at first is unusual and different from the previous practice; prepare the plans to mobile platform first.

This method has two important advantages: one is a business and the other one is a design aspect benefit.

1) *Business benefit: Growing usage of mobile devices*

A November 2013 survey of Business Insider [10] shows that 60 percent of online devices are smartphones or tablets, mobile data traffic has a 20 percent share of the whole Internet traffic and 20 percent of mobile users use their mobile devices for online purchases; moreover all three metrics show clear increasing trend.

2) *Design benefit: Thoughtful content*

A smaller screen size helps to carefully ensure the importance of the order of services, size constraints force us to overshadow the less essential elements. A clear, concise user interface has bigger marketing value, quicker and most importantly: the mobile user is satisfied. Of course, more graphical elements, more pictures, more detailed charts, more textual description can be placed on a larger screen, but if we extend an originally well thought content, the relevant information will always remain in focus.

C. *Iterative design and software development*

We described the development steps of the UI in the previous chapter; the iterative process used there is the part of the complete development process.

Once the graphical design of a user interface has created and ready, a HTML template along with the system plan are transferred to front-end and back-end developers. On this level, software engineering follows the agile method, a software module is created based on the system design, followed by a test and further refining process; this process is also iterative (Fig. 6).

D. *Test, user feedbacks*

Whenever the design and the code have teamed up, the testing phase comes, and then a freshly born newer edition can be shown to the user. Thus, the customer can immediately test and comment the current system and its functionality.

Then further correction, another iteration phase comes based on user feedbacks. It is worth to mention that we are talking about only one or two-week cycles, so the development process is very fast.

The customer is satisfied; both because developers constantly consult with him and on the other hand he can continuously see the growth of his envisioned system.

E. *Beta test, final product*

After all system components are in their correct place, the system is ready for beta testing, quality control, minor bug fixes and then the final product can be presented to the market.

This step obviously cannot be described by two sentences but as it is shown on Fig. 6, the process returns again to the sequential waterfall model, so due to space limitations we ignore the presentation of this stage.

Coordination and management of a whole project requires special project and team management efforts, so compared to the classical software development methodologies, this method requires significantly more time, energy and qualities from the whole developer team.

VI. Conclusion

In our rapidly developing world it is essential to immediately respond to technical development. So true is it in the field of web engineering. New devices, new technologies and even new customer needs appear in every year, every month, every day.

We have to change, improve and upgrade not only our software development environments, but also our existing development, engineering practices, methods. In the age of responsive web, developers need such a dynamic, adaptive new method that we expounded in this article. Although we already use this new model in our real web engineering works, only practical exercises can prove the usability and legitimacy of the adaptive design process.

References

[1] T.A. Powell, *Web design: The complete guide*. New York: McGraw-Hill, 2000.

[2] B. W. Boehm, *A spiral model of software development and enhancement*. Computer, 21(5), 61-72., 1988

[3] J. Hampton, "Prototype models of concept representation.", 1993

[4] J. Conny and C. Bucanac, "The V-Model." IDE, University of Larlskrona/Ronneby 1999.

[5] A. Adamko. (2014, May 10) Modeling Web-based Information Systems [Online]. Available: http://ganymedes.lib.unideb.hu:8080/dea/bitstream/2437/78505/5/Web%20alapú%20Információs%20Rendszerek%20modellezése.pdf

[6] S. Murugesan, A. Ginige, *Web Engineering: Introduction and Perspectives*. Software Engineering, 1999

[7] S. Viljami. (2014, Jan 11) Responsive Workflow [Online]. Available: http://viljamis.com/blog/2012/responsive-workflow/

[8] M. Boulton. (2014, Feb 12) Responsive Summit Workflow [Online]. Available: http://www.markboulton.co.uk/journal/responsive-summit-workflow

[9] UXMovement. (2014, May 05) Why it is important to sketch before you wireframe [Online]. Available: http://uxmovement.com/wireframes/why-its-important-to-sketch-before-you-wireframe/

[10] H. Blodget, T. Danova (2014, Jan 11) The Future of Digital:2013 [Online]. Available: http://www.businessinsider.com/the-future-of-digital-2013-2013-11?op=1

Cellular Automaton pRNG with a Global Loop for Non-Uniform Rule Control

Alexandru Gheolbanoiu, Dan Mocanu, Radu Hobincu, and Lucian Petrica
Politehnica University of Bucharest
alexandru.gheolbanoiu@arh.pub.ro

Abstract—Pseudo-random number generation is an important ingredient of many cryptography applications, as well as scientific applications based on statistical sampling, e.g., Monte Carlo methods. Several methods have been proposed for generating pseudo-random numbers, but these are generally either (i) based on cryptographic cypher algorithms and expensive to implement in hardware (e.g., large silicon area, low energy efficiency) or (ii) based on linear-feedback shift registers, which can be efficiently implemented in hardware but are not sufficiently random. This paper presents a pseudo-random number generator which utilzes a configurable cellular automaton network which generates the output stream of numbers, and a feedback loop which monitors the randomness properties of the output stream and adjusts the parameters of the network in order to optimize its cryptographic performance. We demonstrate that introducing this additional feedback loop increases the overall entropy of the system, improving the quality of the pseudo-random sequence over other cellular implementations or LFSRs. We also analyze the effect of multiple configurations of the proposed generator architecture. We evaluate the generator against several standard benchmarks to illustrate its performance and we also provide an evaluation of its hardware implementation which demonstrates comparable implementation efficiency to LFSRs.

Keywords—cellular automaton, random number generator, LFSR, feedback, FPGA.

I. Introduction

Random number generators (RNGs) are essential for the generation of cryptographic keys for secure online communication, and have become a necessary part of any digital system. Other applications of RNGs are are statistical simulation algorithms based on the Monte-Carlo method and even video games. Since all of these applications can be executed on, e.g., a desktop computer, the processing system needs to include a RNG of sufficiently good quality. A true RNG is desirable because one cannot predict its output under any circumstances, and once a sequence of such numbers has been generated, someone cannot predict if and when it will be generated again [1]. True RNGs are difficult to implement in a digital system because such a system is inherently deterministic, but physical processes like the initialization of random access memories may be utilized for RNG purposes [cite something here].

The easier approach is to combine different algorithms and mathematical functions for the purpose of generating numbers that create the appearance of randomness. These generators are called Pseudo Random Number Generators (PRNG). The most popular such PRNGs are the Linear Feedback Shift Registers (LFSR) due to their efficient hardware implementation [2]. However, these generators present an unwanted property: periodicity. Thus, after a sequence of N numbers, where N is the repetition period of the generator, the exact same sequence will start being generated again. Attempts to increase the period N are made through the use of Non-Linear Feedback Shift Register (NLFSR) [3] [4] [5]. The research on these generators is still ongoing and the construction of large NLFSRs with guaranteed long periods remains an open problem.

Several researchers have attempted to harness the properties of cellular automata for random number generation. A cellular automaton (CA) is a network of cells in a finite dimension space, whereby each cell has a set number of possible states which are updated periodically based on a rule which takes into account the previous state and the states of other cells in a neighbourhood. A CA may be uniform, meaning all cells have the same rule, or non-uniform, with different cells having different rules. The one-dimensional (1D) and two-dimensional (2D) automata have seen more research interest, with the most well-known cellular automaton being Conway's Game of Life [6], a 2D automaton which has been proven to be Turing complete [7]. Much of the research on CA-based RNG has focused on the identification of CA rules which lead to good randomness properties. Wolfram in particular has performed extensive analysis on 1D cellular automata rules, and for the remainder of this work we will utilize the rule naming conventions defined in [8].

In this work, we attempt to create a one-dimensional CA RNG with good randomness properties by analyzing the cell network in its entirety. Instead of integrating different algorithms and hardwired mechanisms within the cells, i.e., searching for the perfect (and most likely complex) CA rule, we maintain the simplicity and flexibility of the cell in order to facilitate hardware implementation. With the addition of a feedback mechanism, which we call the global loop and which is able to reconfigure the cell rules, RNG properties are improved, as demonstrated with industry-standard randomness benchmarks.

This paper is structured as follows. Section II will present existing work on PRNGs, with focus on methods based on cellular automata. Section III introduces the proposed system architecture and Section IV describes its implementation. Section V presents the evaluation methodology and results, while in Section VI we make concluding remarks and outline areas for future research.

II. Cellular Automaton RNGs

In 1986, Stephen Wolfram first tried to construct RNGs through the use of Cellular Automata [9]. His main focus was to demonstrate that one-dimensional uniform CA networks are able to generate random numbers of higher quality than most LFSR generators [10]. Since then, multiple works have attempted to improve on the idea of 1D CA RNGs [11] [12] [13]. Some attempts have even been made on the construction of 2D CA RNGs [14]. Most of these studies focused more on the CA cell itself, the rules to be used and the evolution of set cells, resulting in complex circuits which, in most cases, are impractical.

Wolframs work focused on analyzing the potential of different CA rules for random number generation with the use of a one-dimensional uniform CA network. In order to do this, he applied, in turn, each of the 256 possible rules to all the cells forming the one-dimensional network and, using a suite of randomness tests, measured the potential of each rule to be used in a RNG [10]. His study concluded with a taxonomy of CA rules, consisting of three classes defined by their potential for randomness, class I being the most predictable and class III being the most chaotic, out of which rule 30 best creates the appearance of chaotic behavior. The first problem that remained was that two output streams generated through the use of the same CA rule, but with different initial seeds, presented a strong correlation with each-other in both time and space. The second problem was the periodicity that the generated numbers presented.

In 1989, Hortensius et al. proposed the first non-uniform CA network for random number generation [15]. Instead of having the entire network of cells apply the same rule, two or more rules would be utilized by different cells in the network. In his research, he evaluated a combination of cells with rule 90 and cells with rule 150 within the same network. This configuration reduced the correlation between two output streams and the periodicity of the generated numbers.

In 1999, Tomassini et al. reanalyzed the potential for chaos of the CA rules in uniform networks, but this time, through the use of the DIEHARD evaluation suite [16]. He also pointed out the importance of site spacing and time spacing in the attempt to reduce the correlation of two random bit streams generated with the same CA rule, but with different initial seeds. Unlike the traditional RNG, with site spacing, not all bits generated at one moment by the network are utilized for the output number. And, with time spacing, only bits generated at a certain moment of time are used for the output number. His evaluation concluded that, utilizing site and time spacing, rule 105 presents the most chaotic behavior, followed by 165, 90 and 150. He also introduced the idea that individual cells can improve their randomness quality through genetic algorithms. Each cell would be able to change its rule at the end of a generation cycle depending on the entropy it and its neighbors presented [16]. This concept was termed cellular programming, and focused on the idea of self-evolving non-uniform CA networks, but on a local scale. From this idea, there have been many published researches that focus on the local evolution and control of a CA cell [17] [18].

Other attempts have been made at improving the CA RNGs through the use of traditional genetic evolution algorithms where the network is analyzed in its entirety and modifications are applied to all the cells depending on the results [19] [20]. This practically steps away from the CA network itself, ignores the local loops made between the cells, but applies a global loop, whereby the output of the network is analyzed and, based on different genetic algorithms, modifications are made to the entire network. This type of configuration holds promise for improved randomness properties, but has not been formally evaluated in the existing literature, until now. Our work aims to construct and evaluate such a CA configuration, with focus on both its RNG properties and the efficiency of its hardware implementation.

III. Global Feedback CA RNG

Taking guidance from previous work, we propose to construct a minimalistic CA RNG which is able to satisfy most of the quality requirements that are now placed on RNGs for cryptographic use. Our goal is to design and formally evaluate a system which is capable of outputting a high entropy sequence of numbers whilst maintaining the hardware resource usage to a minimum.

The principal challenge towards the intended goal is how to prevent the CA RNG from remaining stuck in a steady state or in short cycles. This requirement may be achieved in one of two ways, either by implementing a generic algorithm in each cell for rule updates, or by implementing a global feedback loop which is be able to control the entire network by updating rules in individual cells. Because we desire to obtain a small structure, the area and complexity of individual cells must be minimized. Therefore, we choose to implement a global feedback loop to control the rules within the CA network based on certain criteria.

Another design choice is whether he global loop can appoint only one rule to the whole network, resulting in a uniform network, or appoint several different rules to different parts of the network, resulting in a non-uniform network. In order to fully take advantage of a global loop, and in following with the findings of previous work on non-uniform CA, we elect to implement a feedback loop which is able to control the rule for each individual cell within the network. Therefore, the CA network will be generally non-uniform and the cells will be capable of retaining rules applied to them by the external loop.

As previously stated, the global feedback loop will collect the output of the network, analyze its properties and apply the required modifications, i.e., change the rules of different cells individually. This process is illustrated in Figure 1. An important design consideration is the nature of the feedback. Evaluations were carried out on negative feed-back mechanisms, e.g., an external system would calculate the entropy of the network and, in case is decreasing, modify the cells rules in an attempt to revitalize the RNG, but these proved incapable of ensuring good randomness. Similar results were obtained

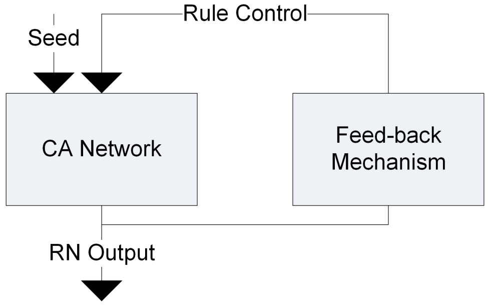


Fig. 1: CA with Global Feedback Loop

utilizing a toggle checker system to count the transitions of each bit of the CA network output within a time frame and, should a bit get stale, call for a rule change. Therefore, negative feedback was discounted and the focus was switched to positive feed-back mechanism, whereby the mechanism collects the generated output of the CA network and, at fixed intervals of time and through the use of the entropy collected, applies a new rule to a randomly selected CA cell.

Another important design consideration is the selection of the new rule to apply to a given cell. Randomly generating a new rule between 0 and 255 was not expected to yield good results, since most CA rules do not exhibit good randomness properties. Additionally, this architecture requires that each cell contain the circuits required to implement all 256 rules, leading to large area footprint and circuit complexity. Therefore, our proposed mechanism chooses between a fixed set of rules. Based on the research done by Tomassini [16], we selected rules 105, 165, 90 and 150 to be the only candidates for a new rule. We decided for all cells to follow rule 105 initially, not only because Tomassinis work proved rule 105 to provide the best randomness, but also because it allows CA oscillation even when the initial state of cell, i.e., the seed, is all zeroes or all ones. For example, if the seed is 0 the CA will oscillate between 0x0000...0000 and 0xFFFFFFFF until the feedback mechanism applies the first rule change, and will subsequently exhibit random output.

Finally, we add to the design a site and time spacing mechanism, as illustrated in Figure 2 in order to avoid the output correlation problems which usually occur with CA RNGs. From the output of the entire CA network, denoted N, of length L_N, the site spacing mechanism selects and passes on only the outputs of cells located at set spatial intervals. The length of the interval is denoted S_s. Consequently, for a site spacing value of 2, only the outputs of cells 0, 3, 6, 9, etc. are utilized for generating an output number. We denote S the output after site spacing, of length L_S. Conversely, if the system is required to generate random numbers of a certain length L_S, with a set site spacing, the required number of CA cells is given by Equation 1. The higher number of cells required increases the CA RNG circuit size, and it is desired to have a site spacing value as small as possible.

$$L_N = (S_s + 1) * L_S \tag{1}$$

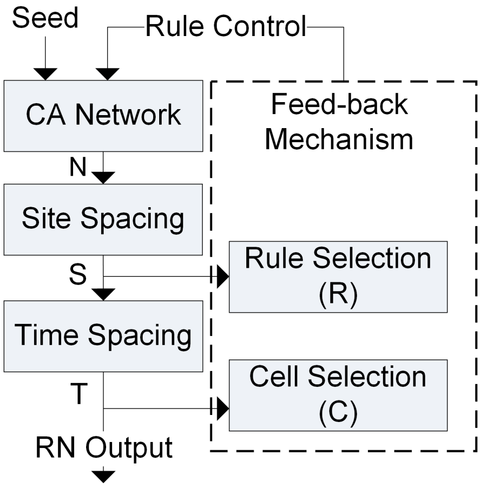


Fig. 2: CA with Site and Time Spacing

The time spacing mechanism forwards, at regular time intervals, the output values of the site spacing mechanism to the output of the CA RNG. For example, if the numbers S_0, S_1, S_2, S_3, S_4, S_5, etc. are output by the site spacing mechanism, with a time spacing value of 1, only S_0, S_2, S_4, etc. are selected. This mechanism decreases the correlation of consecutive output words but reduces the RNG throughput. The output of the time spacing mechanism is denoted T. The time spacing value T_s is also desired to be as small as possible, in order to reduce the number of circuit operations per output word, and therefore increase the circuit energy efficiency.

In order to select one of the four available CA cell rules, the rule selection mechanism collects the output of the site spacing and determines the selection bits R_0 and R_1 according to Equations 2 and 3. These two bits will be generated using the collected randomness of the site spacing output.

$$R_0 = (\sum_{i=0}^{L_{RN}/2-1} S_i + R_0) \mod 2 \tag{2}$$

$$R_1 = (\sum_{i=L_{RN}/2}^{L_{RN}-1} S_i + R_1) \mod 2 \tag{3}$$

At each CA network generation cycle, a new value is output by the site spacing circuit and a new rule is selected. The cell selection block determines when and which cell is to be modified. There are three possibile strategies for the timing of rule changes, (i) at each generation cycle, (ii) at a fixed interval or (iii) at a random interval. Strategy (i) is expected to be inffective, since the cell selection mechanism does not have enough time to gather entropy information, which leads to a poor rule selection randomization. Strategy (iii) is discarded because the system has a single entropy source and both the change interval and cell selection would be calculated based on it, resulting in a strong correlation between the two. Hence, we opted for strategy (ii), updating the cell rules ar fixed intervals of time. When the time to change a rule is reached, the block

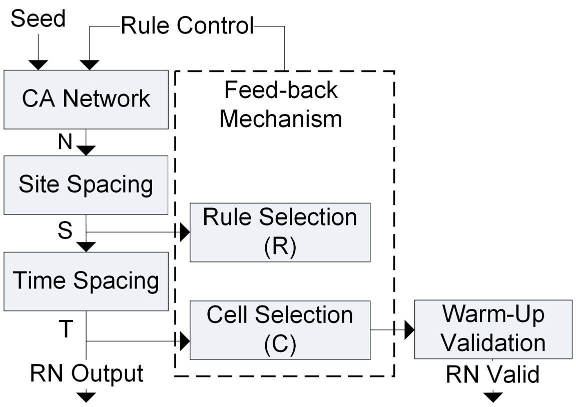


Fig. 3: Final CA RNG Structure

reads the current time spacing output and, based on its value, selects a cell C to have its rule changed with the one currently selected by the rule selection block. This strategy creates a new condition: the rule change interval needs to be at least greater than the time spacing interval, else two consecutive changes may be applied to the same cell only because time spacing has not yet generated a new number.

$$C = T_s \mod L_N \tag{4}$$

The two mechanisms described above ensure a good randomization of the cell rules based only on the output of the network and with minimal circuit complexity and size. We provision an additional external connection to the rule selection block so that the user may input extra random values in order to increase the gathered entropy. The CA network is initially uniform, with rule 105 controlling all the cells, but becomes non-uniform after the first interval of the rule change mechanism in the feedback loop. Consequently, the quality of the first generated numbers will strongly depend on the initial seed. To remove this weakeness in the RNG design, all output numbers of the RNG are discarded until a certain number of rule changes has have occurred. This is called the warm-up period and in order to control it, we introduce an additional block into the design. The warm-up period value represents the number of rule changes that must occur before the output numbers are considered valid. Notably, another reason for not using a random interval for rule changing is that the length of the warm-up period could not be predicted and may in some cases become very large.

Most parameters described above as being part of the different mechanisms (time spacing, rule change interval and warm-up period) are controllable by the user depending on the required performance. The generated number length and site spacing values are constant because they directly impact the number of cells used and the system structure. Tomassini recommends in his work a site spacing of 1 or 2 and a time spacing between 1 and 4. The optimal values for the rest of the parameters are determined after implementation and analysis of the RNG.

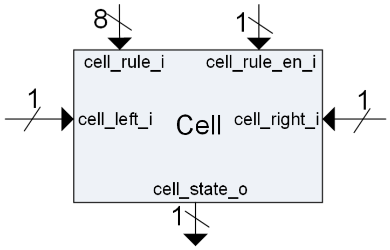


Fig. 4: CA Cell

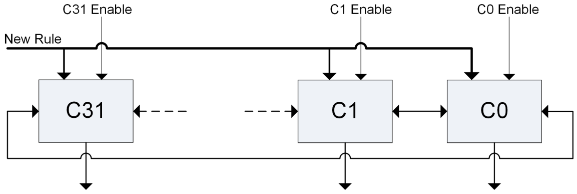


Fig. 5: CA with null site spacing

IV. Implementation

In order to enable the evaluation of the proposed CA RNG hardware structure, we implemented it in VHDL. In our proposed CA RNG architecture, in order for the cells to be able to retain the assigned rule and the current state, each requires 9 flip-flops, 1 for the current state and 8 for the rule storage. Because each cell may have only one out of four rules, two flip-flops for rule storage would be enough to retain a corresponding encoded value. However, we desired to implement a more flexible structure which allows us to experiment with multiple feed-back rule control mechanisms. Hence, each cell requires an additional 8 bit port for the new rule input and an update enable 1 bit port. A diagram of the cell, with all inputs and outputs, is presented in Figure 4.

The CA cells are arranged as a 1D network with its extremity cells sharing a connection, as illustrated in Figure 5. As mentioned in Section III, the number of required cells is given by the site spacing value and the generated number length. For our analysis, we select a length of 32 bits and, with a site spacing of 0 or 1, we require 32 or 64 cells within the network. In order to reduce wiring fan-out issues, new rules are transmitted via a common 8 bit bus to all the cells, while each cell has an individual enable signal.

The site spacing mechanism is implemented by connecting only the appropriate state outputs of the network to the output. The time spacing mechanism contains a counter-based timer which, upon reaching the selected time spacing interval, signals a buffer to store the output of the site spacing mechanism. The output of the buffer is connected directly to the output of the RNG. The cell selection block and warm-up validation block also consist of counter-based timers which signal when the rule change takes place and when the warm-up is done. The rule selection mechanism is illustrated in Figure 6 and consists of a memory for the two rule selection bits, i.e., 2 flip-flops, which are updated according to Equations 2 and 3 at each generation cycle.

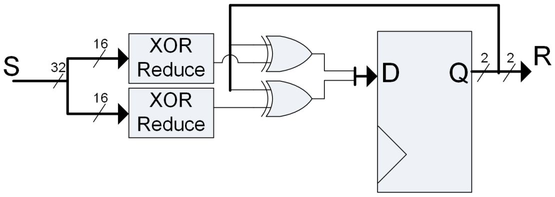


Fig. 6: Feed-Back Rule Selection

The implementation is parametric, with 4 architectural parameters: site spacing, time spacing, warm-up period and rule change interval. We encoded these configurations with a unique name containing all the parameters: S[site spacing]T[time spacing]R[rule change interval]W[warm-up period]. For example S1T3R5W50 is the configuration with site spacing 1, time spacing 3, rule change interval 5, and the warm-up period is 50 rule change intervals.

V. Evaluation

This section presents an evaluation of the proposed CA RNG structure with regard to randomness properties and the efficiency of its hardware implementation on FPGA, i.e., area and maximum frequency.

A. Methodology

We targeted the Xilinx Virtex-6 FPGA architecture for the evaluation of the implementation efficiency, and utilize LUT count and maximum frequency as metrics. To determine whether the system exhibits good randomness properties, we utilize three popular RNG evaluation suites, namely the ENT [21], DIEHARD [22] and NIST [23] suites. Our intention is to verify that the RNG performs well on all the selected suites, of which ENT is the least demanding and the NIST, issued by the foremost authority on public information security in the United States, is the most difficult.

Simulations are performed with site spacing 0 and 1, time spacing 0 to 5, warm up period of 50 and rule change interval of 1 to 5. Utilizing these parameter values and the associated simulation environment we obtained a set of 30 sequences of random numbers, each generated by the CA RNG with a specific configuration. Site spacing beyond 1 is not evaluated because the hardware implementation is expected to become unfeasably large. Of these 30 configurations, we keep only those that do not exhibit short cycles, i.e., an output value is not observed more than once in every seven consecutive output words.

The first evaluation is performed on ENT, which runs 7 tests to help discern the quality of the random sequence. These tests are entropy, optimum compression, chi square distribution, arithmetic mean, Monte Carlo value for Pi, and serial correlation coefficient. ENT is the only benchmark which outputs a set of absolute results for the 7 tests it runs. Although it is not generally regarded as the most relevant benchmark for random number generators, the fact that it outputs absolute values allows us to utilize the results for selecting a number of configurations to go forward. We select the best performing configurations for each test, which continue to the DIEHARD and NIST statistical benchmarks.

TABLE I: Inter-stream correlation

Configuration	Seed 1	Seed 2	Correlation
S1T2RxW50	Binary '0'	Binary '1'	-0.0038
S1T2RxW50	Pattern 0xA	Pattern 0x5	0.001747
S1T3RxW50	Binary '0'	Binary '1'	0.000995
S1T3RxW50	Pattern 0xA	Pattern 0x5	-0.005248

DIEHARD contains 12 statistical tests that output a p-value, which should be uniform on [0,1) if the input file contains truly independent random bits. A p-value of 1 or 0 means the input sequence has failed the test. After validating the remaining CA RNG configuraton with DIEHARD we continue by running the NIST Suite. This evaluation suite has been developed by the Random Number Generation Technical Working Group (RNG-TWG) between 1997 and 2010 as a benchmark for RNGs and PRNGs used in cryptographic applications. NIST contains 15 tests and, similar to DIEHARD, outputs a p-value that determines if the input sequence has passed or failed the test.

Finally, we evaluate the remaining configurations on TestU01 [24], a benchmark consisting of four sub-tests. We remove from our initial set of CA RNG configuration those that have failed one of the randomness benchmarks, and evaluate the remaining configurations for FPGA implementation efficiency. As target FPGA architectures, we select Xilinx Spartan-3 and Virtex-6. Spartan-3 is selected for direct comparison to previous work on FPGA random number generation in [25], while Virtex-6 is a more modern architecture.

B. Results

The initial short cycle evaluation results in the elimination of all configurations with null site spacing, therefore leaving only 15 configurations for further analysis. The ENT evaluation does not further eliminate any of the remaining configurations, as all exhibit good performance on the ENT banchmarks. The NIST benchmark passes on all remaining configurations. DIEHARD fails on all configurations with time spacing smaller than 2, therefore only 10 out of the initial 30 configurations are selected for evaluation with TestU01. Of these, all except S1T3R4W50 pass the randomness test.

We also analyzed the inter-stream correlation of the winning configurations, which is the correlation between streams generated with the same configuration but with different seeds. Ideally, output streams from different seeds are completely uncorrelated. The correlation evaluations were performed between two pairs of seeds, consisting of the binary representation of decimal values 0 and 1, and the repeating patterns of 0xA and 0x5 respectively. The calculated correlation is a number between -1 and 1, ideally 0. The correlation between the streams generated by the seeds are presented in Table I. All configurations with the same rule change interval performed identically and were therefore compressed in the same table entry.

TABLE II: FPGA Implementation

Configuration	Spartan-3 LUT/FF	Spartan-3 F_{max}	Virtex-6 LUT/FF	Virtex-6 F_{max}
S1T2R2W50	358/380	177	237/227	641
S1T2R5W50	358/380	177	241/231	641
[25]	307/202	181	-	-

Finally, we synthesized the circuit for the target FPGA architecture of Xilinx Spartan-3 and Virtex-6. Table II gives an overview of the best and worst implementation results of the evaluated configurations, set against implementation results from previous work in Thomas et al. [25]. From previous work we selected the smallest implementation which passed the DIEHARD and Crush benchmarks, since Crush is a part of TestU01. For all remaining CA configurations, the theoretical maximum frequency is calculated at 600 MHz, and estimated area is similar. It must be noted that, while Spartan-3 results are comparable, our work is optimized for Virtex-6 and therefore Spartan-3 performance may suffer.

VI. Conclusion and Future Work

We have presented a pseudo-random number generator consisting of a one-dimensional cellular automaton and a feedback loop which monitors the CA outputs and modifies the CA rules at set time intervals in order to improve the randomness properties of the RNG. The generator was designed with hardware efficiency in mind, and the resulting structure is capable of passing all the selected randomness benchmarks, while also occupying very little area when implemented in a modern FPGA and is capable of operating at a high frequency of over 600 MHz.

Future work will focus on the continued analysis of the randomness properties of the proposed CA RNG architecture, and on comparisons to other methods of random number generation, with regard to both quality of output stream and hardware implementation efficiency. In this work we have explored a small number of the possible configurations of the CA RNG architecture, and future work will also concentrate on expanding the analysis to a larger number of configurations.

Acknowledgment

Part of this work was carried out with funding and support from POSDRU/159/1.5/S/132397 ExcelDoc postdoctoral program.

References

[1] S. Srinivasan, S. Mathew, R. Ramanarayanan, F. Sheikh, M. Anders, H. Kaul, V. Erraguntla, R. Krishnamurthy, and G. Taylor, "2.4 ghz 7mw all-digital pvt-variation tolerant true random number generator in 45nm cmos," in *VLSI Circuits (VLSIC), 2010 IEEE Symposium on*. IEEE, 2010, pp. 203–204.

[2] S. W. Golomb, L. R. Welch, R. M. Goldstein, and A. W. Hales, *Shift register sequences*. Aegean Park Press Laguna Hills, CA, 1982, vol. 78.

[3] E. Dubrova, "A list of maximum period nlfsrs." *IACR Cryptology ePrint Archive*, vol. 2012, p. 166, 2012.

[4] R. Gottfert and B. M. Gammel, "On the frame length of achterbahn-128/80," in *Information Theory for Wireless Networks, 2007 IEEE Information Theory Workshop on*. IEEE, 2007, pp. 1–5.

[5] B. Gammel, R. Göttfert, and O. Kniffler, "Achterbahn-128/80: Design and analysis," in *ECRYPT Network of Excellence-SASC Workshop Record*, 2007, pp. 152–165.

[6] J. Conway, "The game of life," *Scientific American*, vol. 223, no. 4, p. 4, 1970.

[7] P. Rendell, "A universal turing machine in conway's game of life," in *High Performance Computing and Simulation (HPCS), 2011 International Conference on*. IEEE, 2011, pp. 764–772.

[8] S. Wolfram, "Statistical mechanics of cellular automata," *Reviews of modern physics*, vol. 55, no. 3, p. 601, 1983.

[9] ——, "Cryptography with cellular automata," in *Advances in Cryptology CRYPTO85 Proceedings*. Springer, 1986, pp. 429–432.

[10] ——, "Random sequence generation by cellular automata," *Advances in applied mathematics*, vol. 7, no. 2, pp. 123–169, 1986.

[11] D. De la Guia-Martinez and A. Fuster-Sabater, "Cryptographic design based on cellular automata," in *Information Theory. 1997. Proceedings., 1997 IEEE International Symposium on*. IEEE, 1997, p. 180.

[12] I. Kokolakis, I. Andreadis, and P. Tsalides, "Comparison between cellular automata and linear feedback shift registers based pseudo-random number generators," *Microprocessors and Microsystems*, vol. 20, no. 10, pp. 643–658, 1997.

[13] M. Matsumoto, "Simple cellular automata as pseudorandom m-sequence generators for built-in self-test," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 8, no. 1, pp. 31–42, 1998.

[14] M. Tomassini, M. Sipper, and M. Perrenoud, "On the generation of high-quality random numbers by two-dimensional cellular automata," *Computers, IEEE Transactions on*, vol. 49, no. 10, pp. 1146–1151, 2000.

[15] P. D. Hortensius, R. D. McLeod, and H. C. Card, "Parallel random number generation for vlsi systems using cellular automata," *Computers, IEEE Transactions on*, vol. 38, no. 10, pp. 1466–1473, 1989.

[16] M. Tomassini, M. Sipper, M. Zolla, and M. Perrenoud, "Generating high-quality random numbers in parallel by cellular automata," *Future Generation Computer Systems*, vol. 16, no. 2, pp. 291–305, 1999.

[17] S.-U. Guan and S. Zhang, "Pseudorandom number generation based on controllable cellular automata," *Future Generation Computer Systems*, vol. 20, no. 4, pp. 627–641, 2004.

[18] D. H. Hoe, J. M. Comer, J. C. Cerda, C. D. Martinez, and M. V. Shirvaikar, "Cellular automata-based parallel random number generators using fpgas," *International Journal of Reconfigurable Computing*, vol. 2012, p. 4, 2012.

[19] G. Stefan, "Looking for the lost noise," in *Semiconductor Conference, 1998. CAS'98 Proceedings. 1998 International*, vol. 2. IEEE, 1998, pp. 579–582.

[20] M. Sipper, *Evolution of parallel cellular machines*. Springer Heidelberg, 1997, vol. 4.

[21] J. Walker. (1998) Ent randomness test. [Online]. Available: http://www.fourmilab.ch/random/

[22] G. Marsaglia and W. W. Tsang, "Some difficult-to-pass tests of randomness," *Journal of Statistical Software*, vol. 7, no. 3, pp. 1–9, 2002.

[23] S. Chari, C. Jutla, J. R. Rao, and P. Rohatgi, "A cautionary note regarding evaluation of aes candidates on smart-cards," in *Second Advanced Encryption Standard Candidate Conference*. Citeseer, 1999, pp. 133–147.

[24] P. L'Ecuyer and R. Simard, "Testu01: A c library for empirical testing of random number generators," *ACM Transactions on Mathematical Software (TOMS)*, vol. 33, no. 4, p. 22, 2007.

[25] D. B. Thomas and W. Luk, "High quality uniform random number generation through lut optimised linear recurrences," in *Field-Programmable Technology, 2005. Proceedings. 2005 IEEE International Conference on*. IEEE, 2005, pp. 61–68.

An Integration of Modeling Systems Based on DSM-platform

Lyudmila N. Lyadova, Alexander O. Sukhov, Elena B. Zamyatina

Abstract **—** An approach of using of the DSM-platform MetaLanguage for integration of various modeling systems is presented. This tool allows to design visual domain-specific modeling languages and to create domain models with developed languages. The MetaLanguage system includes components for describing transformations of models from one formal notation to another. Domain-specific modeling permits various specialists to use concepts from different domains at creating and analyzing of models. An integration of DSM-platforms with tools of models analysis allows to involve domain experts, end-users in the process of constructing and analyzing of models; to reduce the complexity of models development; to fulfill research of models from various points of view with usage of various methods and tools.

Keywords **—** Domain-specific languages, language workbench, modeling languages, models transformation, simulation.

I. Introduction

Information and analytical systems, which are used for solution of various management tasks, are created with technologies, which are based on the models. Mainly for models creation graphical notations, diagrams of various types are used. These notations and diagrams allow to describe objects of the modeled business system, their properties, relations between them, operations executed over them, business processes, etc.

The important conditions for reducing the complexity of users work are the possibility of integration of various information systems, the reusing of created models, and their transfer from one system to another for solving of various tasks. The transformation of models from one modeling language to another can be required [4].

These requirements can be implemented on the basis of creation of *domain-specific modeling* (DSM) tools, which are called the *DSM-platforms* (*language workbenches*), the main purpose of these tools is development of high-level *domain-specific languages* (DSLs), designed to create models of systems, focused on solving problems in various domains [1], [2]. The language workbench can become the basis for integration of various tools intended for development of information systems, based on the created models (CASE-tools), and for systems analysis (in particular, simulation systems) [3]. At DSLs usage not only domain singularities, but also qualification of users can be considered.

Maximal flexibility of modeling tools may be obtained at creating the multilevel models describing the modeled systems from various points of view and with different levels of details. For matching of various system descriptions it is necessary to construct the whole hierarchy of models: model, metamodel, meta-metamodel, etc., where *model* is an abstract description of system characteristics that are important from the point of view of the modeling purpose, *metamodel* is a model of the language, which is used for models development, and *meta-metamodel* (*metalanguage*) is a language, on which metamodels are described [5].

As a part of the delivered problem the complex of tasks on creation of DSM-platform, which satisfies the following *requirements*, should be solved:

- possibility of modeling languages constructing for a wide range of domains;
- possibility of multi-level modeling (it allows to modify the metalanguage description, to extend it with new constructions, thus approaching the metalanguage to the specificity of domain);
- possibility of modification of modeling language description without regeneration of source code of DSLs editor;
- automatic support in a consistent state of the metamodels and models description at modification of a metalanguage or a metamodel;
- uniformity of tools of representation, description and usage as models and metamodels: creation of models at different levels of hierarchy and operation with them should be carried out uniformly, using the same tools;
- availability of tools for models transformations that allow to convert models as between different levels of the hierarchy, and within the same level (between various modeling languages);
- usability of language toolkits for various categories of users: professional developers (programmers, system analysts, date base designers, etc.), domain experts, business analysts, end-users.

This work was supported in part by Russian Foundation for Basic Research (grants 14-07-31330, 12-07-00302, 13-07-96506).

E. B. Zamyatina is with the National Research University Higher School of Economics, Perm, Russia (e-mail: E_Zamyatina@mail.ru).

L. N. Lyadova is with the National Research University Higher School of Economics, Perm, Russia (e-mail: LNLyadova@gmail.com).

A. O. Sukhov is with the National Research University Higher School of Economics, Perm, Russia (phone: (+7) 912-589-0986; e-mail: Sukhov.psu@gmail.com).

II. Development of Models in MetaLanguage

All the possibilities and demands mentioned above are not realized in any language workbench nowadays [6]. But DSM-platform MetaLanguage attempts to overcome these disadvantages. The MetaLanguage system is designed to create visual dynamic adaptable domain-specific modeling languages, to construct models using these languages and to transform created models in various textual and graphical notations.

One of the basic elements of language workbench is the metalanguage. The *basic elements* of the metalanguage of the presented system are entity, relationship and constraint [7]. The *entity* describes a particular construction of modeling language, i.e. it is the domain object, important from the point of view of the solving problem. The *relationship* is used for describing a physical or conceptual links between entities. The Metalanguage system allows to create three types of relationships: *association*, *aggregation*, *inheritance*. The *constraints* define the rules of models constructing. The constraints are defined for the entities and relationships between them.

Let’s describe the process of building models using the MetaLanguage system (see Fig. 1).

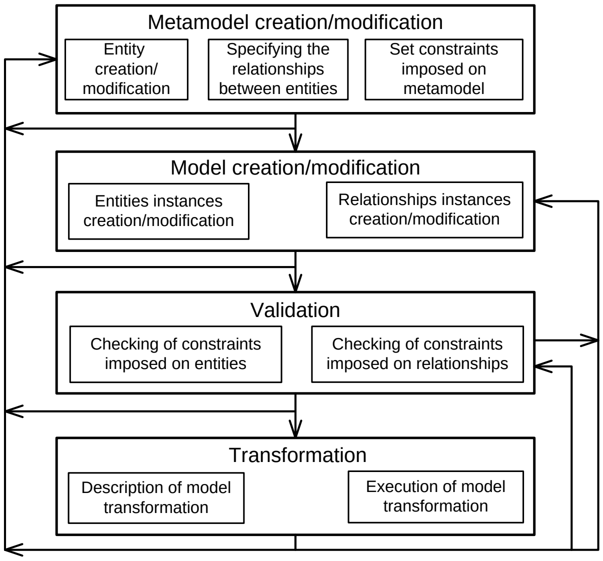

Metamodel creation/modification
Entity creation/ modification
Specifying the relationships between entities
Set constraints imposed on metamodel
Model creation/modification
Entities instances creation/modification
Relationships instances creation/modification
Validation
Checking of constraints imposed on entities
Checking of constraints imposed on relationships
Transformation
Description of model transformation
Execution of model transformation

Fig. 1. Process of creation/modification of domain models with MetaLanguage system usage

The first stage supposes *developing of a metamodel*. The metamodel is a domain-specific language intending to solve specific problems of analyzing domain. Metamodel developers have to use a model’s editor. The developer obtains an extensible dynamically customizable visual modeling language as a result of metamodel creation.

Then users (model designers, system analysts) can *develop models*, which contain instances of specific entities and relations between them, with application of constructions of created DSL. Thereafter, it is necessary to *validate the developed model*: to check if all constraints for the entities and relations between them are met.

A developer can *store designed models in repository*.

User can *transform model* in accordance to rules defined in system [8]. So the designed model can be translated to one or other languages and can be exported to external program systems (simulation system, CASE-tool, for example).

Developed DSL can be used as a metalanguage. The whole *hierarchy of languages* can be created on its basis. This hierarchy allows to work with models of various abstraction levels, focused on solving of various tasks by different categories of users in terms of their domain [9]. At metamodel modification, the system automatically will make all necessary changes in models created on the basis of this metamodel.

Different categories of users are involved in development of domain-specific languages and creation of models with their usage (see Fig. 2).

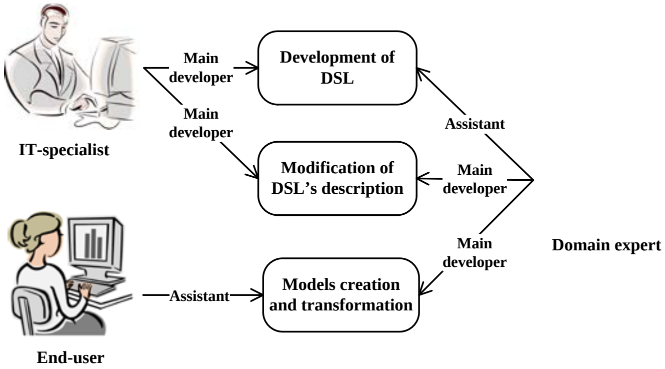


Fig. 2. Involvement of various categories of users in the development of DSLs and models

So developers (IT specialists) with the direct participation of domain experts create DSL, describing the basic concepts of the domain, relationships between them and the constraints, imposed on a metamodel, define rules of models transformations. Domain experts and end-users with the developed language build domain models and fulfill transformations. If it is necessary to modify modeling language the domain experts can independently make appropriate changes in language description or invite the IT-specialists for performing of all necessary modifications.

III. DSM-platform as a Base of Integration

Creation of information systems with usage of modern CASE-tools is based on development of the various models describing the information system domain, defining data structures and algorithms of system functioning [10]–[13]. The choice of tools frequently determines also a choice of language for models description. Thus, the used tool actually “imposes” to developers and users a specific modeling language, which more often operates with terms of some programming paradigm, therefore tools do not allow to domain experts to participate in development and modification of models, that is a necessary condition for creation of effective management systems, increase of efficiency of their adaptation, reducing of maintenance complexity.

One of approaches to solving of this problem is integration of DSM-platforms with tools of information systems development or directly with the information systems, which fulfill interpretation of models at the stage of functioning. Thus, the DSM-platform can become the basis for integration of various tools intended for development of the information systems on the basis of created models and for the analysis of systems via formal models. So, for example, the MetaLanguage system can be integrated with CASE-tools, business analysis systems, simulation systems (see Fig. 3).

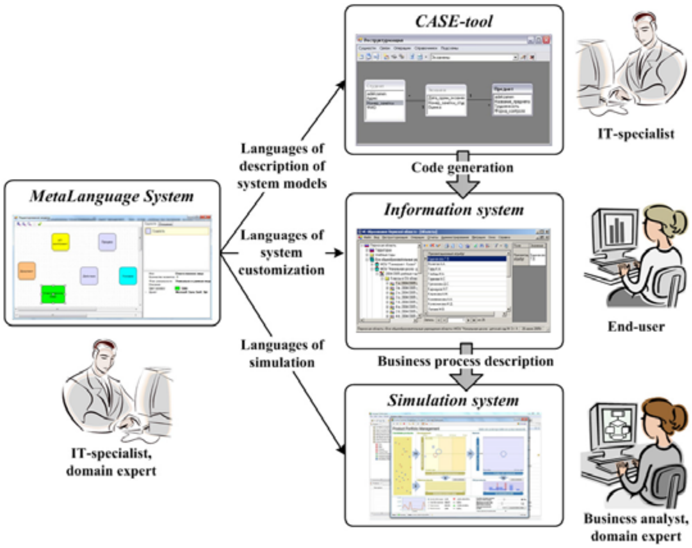

CASE-tool
IT-specialist
Languages of description of system models
Code generation
MetaLanguage System
Information system
Languages of system customization
End-user
Languages of simulation
Business process description
Simulation system
IT-specialist, domain expert
Business analyst, domain expert

Fig. 3. Scheme of usage of MetaLanguage system in the process of creation and maintenance of information systems

Instead of describing models in the notation of visual general-purpose languages, experts with usage of MetaLanguage system can develop DSLs for creation and maintenance of models. After designing of domain-specific languages, the developers with the participation of domain experts create models of information systems. For export of designed models to CASE-tool, it is necessary preliminarily to fulfill conversion of model description to one of standard modeling languages, supported by this tool.

If the developed product is an information system with interpretation of models, the DSM-platform can be used even at the stage of system functioning, so end-users can modify description of domain model, business processes developed with usage of DSLs to adapt system for new conditions.

For business processes analysis and optimization the simulation systems can be used. For carrying out simulation experiments, it is necessary to transform business processes models, created with usage of DSLs or with notation of other modeling languages, in graphical/textual notation, supported by simulation system. After research of business processes their reengineering with usage of domain-specific languages, created in the MetaLanguage system, without source code regeneration and participation of IT-experts can be fulfilled.

Thus, the MetaLanguage system can be used both at the process of development and maintenance of information systems, and as the extension of systems analysis tools.

IV. Integration of MetaLanguage System with Simulation Systems

It is known that in some cases, simulation is a single method of research of complex dynamic systems, and it is widely applied in various fields of science and industry. Development of science and technologies, and, hence, increase of complexity of researched systems, puts more and more complex tasks for simulation. For obtaining of reliable information during simulation experiments it is required to involve experts from different fields of knowledge, and therefore, simulation software should allow to researchers to work in various modeling environments, using different systems of concepts, varied visual or textual languages. For example, at business processes modeling, researchers can attract graph theory, Petri nets or queuing networks. In this case, it is necessary, that a modeling system submitted the user possibility of using not only of various mathematical apparatus, but also of various modeling languages, which operate with terms clear to various categories of users.

Queuing networks (QN) are widely used to analyze the characteristics of business systems in various areas. Various methods and tools (statistical analysis, simulation, etc.) are used for research. Let's consider the example: we'll design new DSL for *QN modeling* with MetaLanguage system and then we'll define rules for the model transformation from designed DSL to the GPSS modeling language.

The *metamodel* of this domain-specific language contains following *entities* (see Fig. 4):

– *Generator* is the entity, which is responsible for generation of requests flow (*transacts* flow), expecting service in system. Intervals between requests arrivals are the random values with a certain distribution. This entity has the following attributes "Name", "Initial delay", "Amount of transactions", "Priority".
– *Queue* is the entity, representing set of transacts, which expect service. It is waiting buffer of the servicing device if it is occupied. The entity "Queue" has the following attributes: "Name", "Maximum length" and "Current length".
– *Servicing device* is the entity, which is responsible for service of requests. The device possesses limited possibilities of transacts service. Handling of request takes some time. The service time is a random value with a certain distribution function. The attributes of the entity are "Name", "Amount of channels", "Service time".
– *Separator* is the entity, allowing to create multiple copies of transacts, each of which will request claim of service. The attributes of this entity are "Name", "Amount of copies", "Block" (name of the block, to which it is necessary to transmit copy of request for service).
– *Collector* is the entity, allowing to integrate multiple transacts flows into a single flow. The entity "Collector" has the attributes "Name" and "Amount of flows".
– *Terminator* is the entity, deleting transacts from model.

- *Distribution* is the entity, which is parent for the entities "Normal distribution", "Uniform distribution", "Student's distribution", etc.
- *Normal distribution* is a distribution, according to which a generation of new requests and/or their service is fulfilled. This entity has two attributes "Expected value", "Variance".
- *Uniform distribution* is a distribution, according to which generation of new requests and/or their service is fulfilled. This entity has two attributes "Minimum value", "Maximum value".
- *Student's distribution* is a distribution, according to which generation of new requests and/or their service is fulfilled. This entity has attribute "Amount of degrees of freedoms".

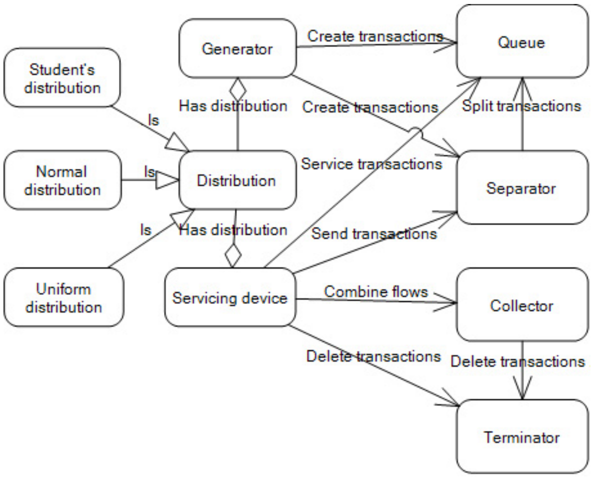


Fig. 4. Simplified metamodel of language for simulation models description

Further, let's describe *relationships* between metamodel entities. As can be seen from Fig. 4, the metamodel contains the following *relationships of association*:

- unidirectional relationship "Create transactions", connecting the entity "Generator" with entities "Queue" and "Separator", shows that after creation of requests they can be placed in a queue for service or be split into multiple flows;
- bidirectional relationship "Service transactions" allows to indicate, what device handles requests in a queue and where they go after service;
- unidirectional relationship "Send transactions", connecting entities "Servicing device" and "Separator", allows to split requests into multiple flows;
- unidirectional relationship "Combine flows" connects entities "Servicing device" and "Collector" and indicates, what collector combines flows of requests for service after their handling by multiple servicing devices;
- unidirectional relationship "Split transactions", connecting entity "Separator" and "Queue", allows to indicate in what queues requests after their separation into several flows should be placed;
- unidirectional relationship "Delete transactions", connecting entities "Servicing device" and "Collector" with entity "Terminator", allows to indicate that after service or combine the transacts should be removed.

The metamodel of modeling language for creation of QN-models also contains three inheritance "Is" that connect the abstract entity "Distribution" with child entities "Normal distribution", "Uniform distribution", "Student's distribution". Child entities inherit all parent entity's relationships.

The aggregation "Has distribution" allows to specify distribution, according to which generation of new requests (transacts) and/or their service is fulfilled.

In Fig. 5 the example of model, constructed with usage of the developed modeling language, is presented. As can be seen from the figure, the model contains generator, four servicing devices (SD1, SD2, SD3, SD4), four queues (QQ1, QQ2, QQ3, QQ4), separator, collector and terminator; two distributions is used. This model describes QN for any domain.

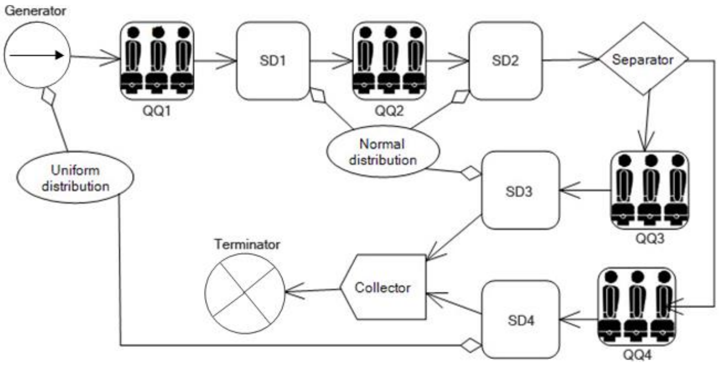
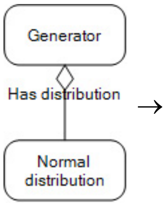


Fig. 5. Examle of QN-model

Let's describe the transformation rules of the constructed modeling language to the notation of GPSS language. User can generate the program in GPSS language applying these rules to the constructed model, and then user can make the analysis of model with usage of the simulation system GPSS.

The rule "Generator_Norm", which converts an instance of entity "Generator", connected by an instance of aggregation relationship with an instance of entity "Normal distribution", into the appropriate command of GPSS language, looks like:

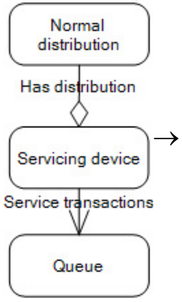


```
GENERATE
   (NORMAL(1,
   <<Normal distribution.Expected value>>,
   <<Normal distribution.Variance>>)), ,
   <<Generator.Initial delay>>,
   <<Generator.Amount of transactions>>,
   <<Generator.Priority>>
```

Symbols "<<" (double opening angle brackets) and ">>" (double closing angle brackets) are used for selection of rule dynamic part, which allows to get values of attributes of entities and relationships instances.

Rules for other types of distributions are described similarly.

The transformation rule "Queue", which converts connected instances of entities "Queue", "Servicing device", "Normal distribution" into the appropriate code of GPSS language, has the following form:



```
→ QUEUE    <<Queue.Name>>
  SEIZE    <<Servicing device.Name>>
  DEPART   <<Queue.Name>>
  ADVANCE  (NORMAL(1, <<Normal
           distribution.Expected value>>,
           <<Normal distribution.Variance>>))
  RELEASE  <<Servicing device.Name>>
```

The rule "Separator" transforms an instance of entity "Separator" into SPLIT command of GPSS language. This rule looks like:

```
→ SPLIT  <<Separator.Amount of copies>>,
         <<Separator.Block>>
```

The rule "Collector", which converts an instance of entity "Collector" into ASSEMBLE command of GPSS language, has the following form:

```
→ ASSEMBLE  <<Collector.Amount of flows>>
```

The rule "Terminator", which converts an instance of entity "Terminator" into the appropriate command of GPSS language, looks like:

```
→ TERMINATE  1
```

After applying of the described transformations to the model presented in Fig. 5 the MetaLanguage system has generated the following code in GPSS language:

```
GENERATE (UNIFORM(1, 2, 8)),,20,100,1
  QUEUE    QQ1
  SEIZE    SD1
  DEPART   QQ1
  ADVANCE (NORMAL(1, 3, 1))
  RELEASE SD1
  QUEUE    QQ2
  SEIZE    SD2
  DEPART   QQ2
  ADVANCE (NORMAL(1, 3, 1))
  RELEASE  SD2
  SPLIT    1, QQ4
  QUEUE    QQ3
  SEIZE    SD3
  DEPART   QQ3
  ADVANCE (NORMAL(1, 3, 1))
  RELEASE  SD3
  QUEUE    QQ4
  SEIZE    SD4
  DEPART   QQ4
  ADVANCE (UNIFORM(1, 2, 8))
  RELEASE  SD4
  ASSEMBLE 2
TERMINATE  1
```

The generated code of model was used for simulation running in GPSS system. The translation of any other visual model developed with created DSL, won't demand additional efforts of model designer or programmer.

V. Conclusion

The Metalanguage system, including transformation component, supports integration of different modeling systems. It provides interoperability of the languages and models in different information and analytical systems. This DSM-platform allows to reduce the complexity of analysts work, to increase efficiency of information systems functioning. Presented language workbench is quite convenient and flexible tool for building of modeling languages and transformation rules of the created models. Usage of the system allows to create DSLs and to determine transformations operatively. Users don't need programming language to develop languages or models. They operate with visual constructions or textual code of initial and target modeling languages.

The research prototype of MetaLanguage system has been used for development of several domain-specific languages, in particular, language for modeling of administrative regulations [14], for manufacturing processes modeling, applications for mobile devices, etc.

References

[1] J. Karna, J.-P. Tolvanen, S. Kelly, "Evaluating the use of domain-specific modeling in practice", in *Proc. of the 9th Workshop on Domain-Specific Modeling at OOPSLA*, Orlando, 2009, pp. 147–153.

[2] M. Velter. (March 2011). MD*/DSL best practices Update March 2011. [Online]. Available: http://www.voelter.de/data/pub/DSLBestPractices-2011Update.pdf.

[3] K. Balasubramanian, D. C. Schmidt, Z. Molnar, A. Ledeczi, "Component-based system integration via (meta)model composition", in *Proc. of the 14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems*, Tucson, Arizona, 2007, pp. 93–102.

[4] S. Sendall, W. Kozaczynski, "Model transformation: the heart and soul of model-driven software development", *IEEE Software*, vol. 20, pp. 42–45, 2003.

[5] J. M. Alvarez, A. Evans, P. Sammut, "Mapping between levels in the metamodel architecture", in *Proc. of the 4th International Conference on The Unified Modeling Language, Modeling Languages, Concepts, and Tools*, Toronto, 2001, pp. 34–46.

[6] A. O. Sukhov, "Comparing of the system of visual domain-specific languages development", *Mathematics of Program Systems*, Vol. 9, pp. 84-111, 2012. (in Russian)

[7] A. O. Sukhov, L. N. Lyadova, "MetaLanguage: a tool for creating visual domain-specific modeling languages", in *Proc. of the 6th Spring/Summer Young Researchers' Colloquium on Software Engineering*, Perm, Russia, 2012, pp. 42–53.

[8] A. O. Sukhov, L. N. Lyadova, "Horizontal transformations of visual models in MetaLanguage system", in *Proc. of the 7th Spring/Summer Young Researchers' Colloquium on Software Engineering*, Kazan, Russia, 2013, pp. 31–40.

[9] E. B. Zamyatina, L. N. Lyadova, A. O. Sukhov, "Multilanguage modeling with MetaLanguage DSM-platform usage", *Informatization and Communication*, no. 5, pp. 11-14, 2013. (in Russian)

[10] K. Balasubramanian, A. Gokhale, G. Karsai, J. Sztipanovits, E. Neema, "Developing applications using model-driven design environments", *Computer*, vol. 39, pp. 33–40, 2006.

[11] J.-M. Favre, "Towards a basic theory to model driven engineering", in *Proc. of the Workshop on Software Model Engineering*, Lisboa, 2004, pp. 48–55.

[12] R. France, B. Rumpe, "Model-driven development of complex software: a research roadmap", in *Proc. of the Workshop on the Future of Software Engineering*, Washington, 2007, pp. 37–54.

[13] J. Hutchinson, M. Rouncefield, J. Whittle, "Model driven engineering practices in industry", in *Proc. of the 33rd International Conference on Software Engineering*, Waikiki, USA, 2011, pp. 633–642.

[14] L. N. Lyadova, A. O. Sukhov, "Modeling of administrative regulations using opportunities of MetaLanguage language workbench", in *Proc. International Conference on Information Systems Development Technologies*, Gelendzhik, 2013, part 2, pp. 45–49. (in Russian)

Security Architecture for Satellite Services over Future Heterogeneous Networks

Vahid Heydari Fami, Haitham Cruickshank
Centre for Communication Systems Research
University of Surrey
Guildford, United Kingdom
{V.Fami; H.Cruickshank }@surrey.ac.uk

Martin Moseley
EADS Astrium
Portsmouth, United Kingdom
{Martin.Moseley}@astrium.eads.net

***Abstract*—The rapid growth in the demand for Future Internet services and many new applications has driven the development of satellite, which are the preferred delivery mechanism due to its wide area coverage, multicasting capacity and speed to deliver affordable future services. However, security has been one of the barriers for satellite services, especially for domains spanning over heterogeneous networks. In this paper, scalable and adaptable security architecture is specified to protect satellite services. The focus is on key management and policy provisioning. Also three scenarios, mobile network, fixed network and Delay Tolerant Network (DTN), are presented, with details on characteristics and security features.**

Keywords - satellite; Heterogeneous network; policy; key management; mobile network; Delay Tolerant Network.

I. Introduction

Satellites will be an integral part of the Future Internet and next generation access technologies such as wireless, mobile and terrestrial broadband. As such, the broadcast nature of satellite coverage can be exploited, costs can be shared among large group of terminals providing a low-cost wide-area Internet multicast service. In addition, group-oriented applications are increasingly deployed over the Internet such as video conferencing, video on demand (VoD), TV over Internet and broadcasting stock quotes. A difficult barrier that prevents the wide exploitation of satellites and the group-oriented applications is the security provisioning for a large and cryptographically heterogeneous multicast group that span multiple domains.

This paper proposes a scalable and adaptable security architecture that protects multicast data according to the cryptographic requirements of a variety of cryptographically heterogeneous domains. This work defines a new satellite multicast security architecture that addresses the specific obstacle that currently impedes development of large scale multicast security services that spans several cryptographically heterogeneous domains. By introducing scalable key management and security policy mechanism, some of the security barriers that inhibit the integration of satellite networks with other network such as next generation mobile networks, would be removed. The major research issues in the security architecture are presented and analyzed, namely key management and security policy provisioning. Also, three sample scenarios are presented, including characteristics and security requirements for each of them.

II. Objectives

Future Internet will be a conglomerate of heterogeneous networks and systems such as satellite, next generation mobile, mobile adhoc and sensors nodes. It is envisaged that satellites will be part of broadband, mobile and Delay Tolerant Networking (DTN) service scenarios, which can span multiple security domains that are cryptographically heterogeneous. The concept of domains is used widely in the Internet. It is also applicable to group-key management to effect scalability, where members are divided (logically or physically) into domains or subgroups. In summary, at least two general types of domains are possible for secure group management:

- Domains according to data encryption: Here, the domains demarcate regions within which differing Traffic Encryption Key (TEK) are used to encrypt the group data.
- Domains according to key management: Here, the domains demarcate key management regions, where each region is associated with a different set of Key Encryption Keys (KEK) for the purpose of managing and disseminating the TEK, which is a common group data key.

Securing such service scenarios could be very challenging due to trust issues, key distribution, policy dissemination/management and multiple encryption/decryption across these domains.

The objective of the work is to specify a scalable and adaptable security architecture that is hierarchical and distributed, in order to protect unicast, multicast and broadcast data for a variety of cryptographically heterogeneous networks. The security architecture involves scalable key management and policy management entities. Such architecture should fit all the three scenarios mentioned above: mobile broadband, fixed network terminal and DTN.

III. System Architecture

This section presents an innovative architecture for securing multicast services across heterogeneous security domains.

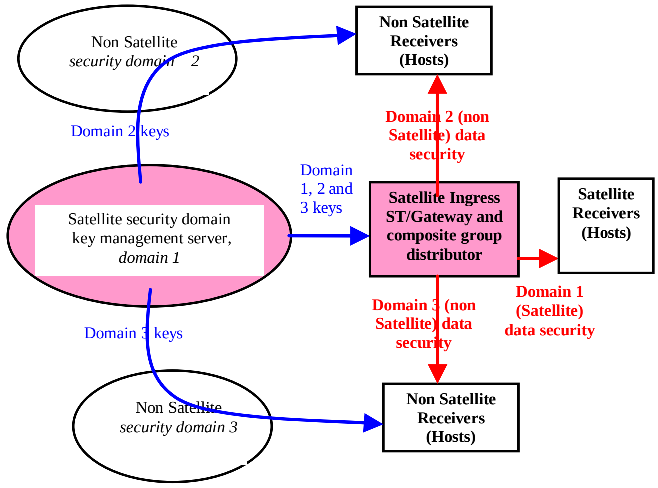

Non Satellite security domain 2
Non Satellite Receivers (Hosts)
Domain 2 keys
Domain 2 (non Satellite) data security
Domain 1, 2 and 3 keys
Satellite security domain key management server, domain 1
Satellite Ingress ST/Gateway and composite group distributor
Satellite Receivers (Hosts)
Domain 1 (Satellite) data security
Domain 3 keys
Domain 3 (non Satellite) data security
Non Satellite security domain 3
Non Satellite Receivers (Hosts)

Figure 1. Secure multicast service across heterogeneous security domains

The architecture for securing multicast services across heterogeneous security domains is shown in Figure1. There are three novel concepts in this architecture:

- The first concept is the adaptive and scalable group key management. It will use adaptive grouping of members into encryption domains (subgroup) that use the same TEK. The partitioning will be made in a way that reduces both re-keying using KEKs and key translation overheads within the overall heterogeneous group. This concept promotes adaptability to changing membership dynamics in various domains.
- The second concept is the use of Data Distributors that disseminate the encrypted data with different keys for each domain. This will eliminate the need for encryption/decryption at security gateways at the ingress of each domain.
- The third concept is the use of security policies, especially for the distributed architecture to delegate trust and role to various entities in each domain. This will promote scalability and adaptability to changing security and threats situations. As such, policies can govern key dissemination, access control, re-keying of group-shared keys, and for the actions taken when certain keys are compromised.

The solution complements the existing link layer security solutions in satellite, digital video broadcasting (DVB), LTE and WiMAX networks. However, it requires that data security should be implemented in a layer in the protocol stack that is common to all domains (e.g. satellite, LTE and WiMAX), such as:

- IP network layer security (using IPSec);
- Transport layer security (TLS);
- Any application layer security.

IV. MAJOR RESEARCH ISSUES

There are several obstacles against the widespread deployment of multicasting services [1][2]. One of them is security. The security mechanisms for unicast are not adequate for the multicast scenario since multicast security mechanisms have scalability and efficiency constraints [3][4][5]. The work proposed in this paper aims to address gaps in secure multicast such as IP Multicast group key management and policies, with a particular focus on a group that spans many domains including a satellite network. Thus, there are two major research issues:

- Multicast key management in cryptographically heterogeneous domains
- IP multicast security policy provisioning

A. Key Management

In a simple case, symmetric cryptography is used by the sender/source and the receivers/destinations, where the data is encrypted by the sender and decrypted by the receivers. The shared key is commonly referred to as the group-key or TEK, since only members of the multicast group are in possession of the key. The use of cryptography necessitates the delivery or dissemination of group keys. Group-oriented security, and more specifically the key management, has been researched for more than two decades. Most of the earlier work has focused on cryptographic approaches to manage keys for hierarchical organizations [6][7][8]. And satellite networks had their research on large scale secure multicast [9][10][11].

Rekeying in secure multicast is needed to preserve forward and backward secrecy whenever members join or leave. Thus rekeying overheads increases as the multicast group gets bigger. The concept of domains is also applicable to group-key management to effect scalability, where members can be divided (logically or physically) into domains (subgroups) [3][12]. However, a clash exists between re-keying overhead and computation overhead for key translation. Finding a trade-off between these two conflicting overheads is essential in the case of networks with resource constrained devices, such as sensors and Mobile Ad hoc NETworks (MANETs) and in the case of very large groups such as satellite multicast. At least two general types of domains are possible for secure multicast management:

- Domains according to data encryption: Here, the domains demarcate regions within which differing group-keys (data keys) are used to encrypt the multicast data. Thus, each domain is associated with a unique group-key, and "crypto-translations" (decryption using one key, followed by encryption using another key) must be carried out at the domain boundaries. Group-members residing within each domain would be in possession of a unique group-key (per domain).
- Domains according to key management: Here, the domains demarcate key management regions, where each region is associated with a different set of key management keys (KM-keys) for the purpose of

disseminating the common group-key (TEK). Thus, each domain would manage its own km-keys (e.g., different rekey period for KM-keys), even though these are used to create safe passage for the common (group-wide) TEK from a key-source, such as a key server, to each of the receivers residing in differing key management domains.

There exist a clash between re-keying overhead, and computation overhead: on one hand, using a single encryption domain increases the re-keying overhead and hence does not scale to large and highly dynamic groups, while it saves computation power which would have been spent in key translation. On the other hand, partitioning the group into different encryption areas reduces the re-keying overhead, but introduces additional computation overhead and delivery delays because of the requirement of key translation. The scalable key management scheme aims to find a good trade-off between these two conflicting overheads.

B. Security Policy

Security policies provisioning is another focal point of the proposed architecture. Similar to other aspects of networking, the correct definition, implementation and maintenance of policies governing the various aspects of multicast security are important factors. Those which are directly related to multicast security include the policies for key dissemination, access control, re-keying of group-shared keys, and for the actions taken when certain keys are compromised [13]. The trust model is a critical issue for secure group communications, which can be established and managed using rule-based security policies. For large scale groups that span several security domains, security management might be delegated to group controllers (key managers) in each domain. Delegation of trust using policies allows the efficient working of distributed security management architecture [14][15]. Thus the use of such policies will help the security integration of satellite network with other networks. Through policies, a system may address the needs of all group participants in real time. The security policy could address the following requirements [16]:

- Identification - Each participant and group can be unambiguously identified.
- Authorization - A group policy can identify the entities allowed to perform protected actions. Group authorization partially determines the trust embodied by the group.
- Access control - Allowable access to group action can be stated by policy.
- Mechanism - Each policy can state how the security requirements of the group are to be addressed.
- Verification - Each policy can present evidence of its validity such as proof of its origin and integrity.

A Reference Framework has been defined and standardized and it addresses all problem areas mentioned above [12][17]. The framework presents a set of functional building blocks that should be tackled for any secure multicast architecture design. It also expresses the complex multicast security from the perspective of architecture (centralized/distributed), multicast group types (1-to-N and M-to-N), and classes of protocols (the exchanged messages) needed to secure multicast packets.

However, currently very little work exists on using security policies for distributed key management, particularly for satellite networks. As such, security policies should be used to delegate trust to key managers and data distributors in various domains. If the multicast group membership is highly dynamic, then policies will also enable adaptive formation and deletion of data encryption domains depending on the subgroup membership dynamics. Security policies are used in the proposed architecture to promote scalability and adaptability in large heterogeneous multicast groups.

V. SCENARIOS

In this section, three scenarios are defined: mobile network scenario for the applications such as mobile broadband, fix network scenario for the applications such as SMART METER, broadband access and Delay Tolerant Network (DTN) scenario for the space applications such as Deep Space. The scenarios are described and the features are discussed in this section.

A. Mobile Scenario

One typical application of mobile scenario is mobile broadband service, which includes web browsing and possibly video streams. Security, as one of the important features of mobile broadband, must be provided to essential signalling messages, but might not necessarily to the large amount of packet data.

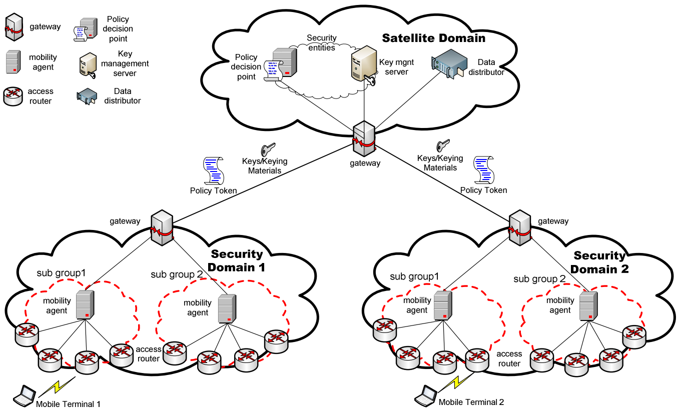


Figure 2. Mobile scenario

As shown in Figure 2, three domains are involved in the mobile scenario: satellite domain, security domain 1 & 2. Security domain 1 &2 are assumed to be cryptographically separated. It is possible that different encryption/decryption algorithm, different key size are used to secure the signalling/traffic data in security domain 1 & 2. The satellite domain provides the ability for centralized key management and policy generation.

1) Satellite domain:

a) Data distributor:

Disseminate the encrypted data with different data keys for each domain. This entity eliminates the need for encryption/decryption at security gateways at the ingress of each domain.

b) Key management server:

Dynamically generate different set of key management keys for different regions. The adaptive and scalable group key management is enabled by the use of key server. It uses adaptive grouping of member into encryption domains that use the same data key, therefore, it reduces re-keying and key translation overheads.

c) Policy decision point (PDP):

It acts as policy server which generates policy (such as policy token [18]) to delegate trust and defines different security mechanisms to various domains. Policy enables adaptable security solutions for changing security and threats situations. Therefore, the resilience to changing security environment is improved. Generally, policy can define key/keying materials dissemination, access control, re-keying conditions, actions taken when a key is compromised, and etc.

It should be noted that the centralized scenario is illustrated in Figure 2. In a centralised scenario, the policy decision point and key server are located in the satellite domain, and relevant security information is disseminated to various security domains. The policy enforcement point (PEP), which cooperates with PDP to enforce policy to the end terminals, can be collocated with entity in each security domain, such as the mobility agent. The PEP can issue policy request on behalf of the end user and handle policy response from the PDP. If distributed system is required, the PDP/key server should be available in each of the security domains, providing the ability to generate policy and set of keys locally within the particular security domain. And the local PDP/key server should be able to operate in a cooperative manner to achieve optimized performance.

2) Security domain1 & 2:

In both of security domain 1& 2, the following entities are involved:

a) Gateway:

It is the point of entry or exist for the security domain, providing connectivity to the satellite domain.

b) Mobility agent:

It provides mobility management service to the mobile terminals, including location updates, forwarding traffic data, and etc.

c) Access router:

It is a layer-3 router, providing network access to the mobile terminal. The access routers can be managed by the mobility agent.

d) Mobile terminal:

It is the mobile user, who would like to use the network resources. It can perform micro-mobility handover within one mobility agent subgroup and can also perform macro-mobility handover across mobility agents/networks.

3) Characteristics:

Some characteristics of mobile scenario are:

a) Moderate bandwidth availability

b) Limited number of security domains

c) Limited coverage areas

4) Security features:

Some security features of mobile scenario are:

a) Specific key management requirements: multiple encryption/decryption domains are needed

b) Moderate data key updates due to moderate data rate in the forward link

c) For multicast services, moderate/fast changing group membership due to the nature of mobile services

d) Either centralized or distributed key/policy management architecture can be considered.

e) For the delay sensitive data in mobile applications, it is required to reduce the negative impact of security on delays by integrating security design with mobility protocols.

f) Due to the nature of valuable bandwidth resources, minimizing signalling overhead introduced by security mechanism is essential. The tradeoffs between strong security design which desired by the cryptography fans and the overhead introduced by security need to be considered.

B. Fixed Network Scenario

The fixed network scenario can be applied to broadband access in rural area, where DSL lines are not applicable, or specific application such as SMART METER/GRID.

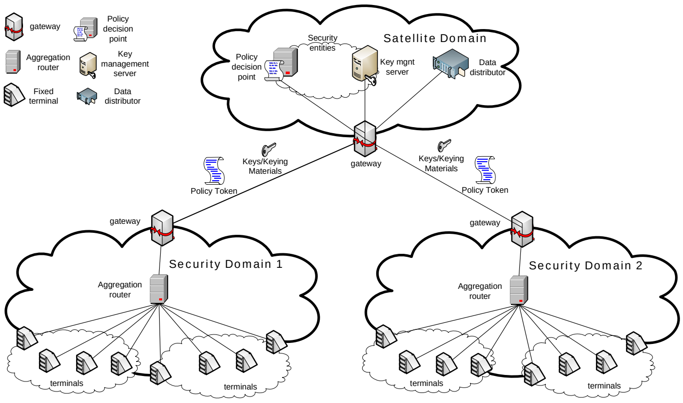


Figure 3. Fixed network scenario

As shown in Figure 3, three domains are involved in the fixed network scenario: satellite domain, security domain 1 & 2. Security domain 1 &2 are assumed to be cryptographically separated. The satellite domain remains the same as in the mobile scenario. While in each of security domain, instead of roaming mobile terminals, there are fixed terminals. The terminals can be broadband service terminals, or other devices, such as SMART METER device installed in the end users' home/office. All of the terminals are connected to the aggregation router, which provides the ability of data aggregation. And the aggregation router is connected to the external network, via gateway. The fixed terminals (for a broadband service) can be connected directly to the aggregation router for the broadband service, and the aggregation router then connects to the satellite access gateway. If the SMART METER application is considered, the terminals are SMART METER devices installed at the end user's home/office to collect the electricity/gas/water meter information. Of all Smart Meter technologies, one critical technical problem is communication. Each meter, especially the sensitive user ID or billing related information, must be reliably and securely transferred to the central location. Considering the varying environments and locations where meters are found, that problem can be daunting. The existing solutions proposed are: the use of cell/pager networks, satellite, licensed radio, combination licensed and unlicensed radio, power line communication (PLC). Not only the medium used for communication purposes but the type of network used is also critical. Fixed wireless, mesh network or a combination of the two have been deployed for SMART METER application. There are several other potential network configurations possible, including the use of Wi-Fi and other internet related networks. No one solution seems to be optimal for all applications. Rural utilities have very different communication requirements from urban utilities or utilities located in difficult locations such as mountainous regions or areas ill-served by wireless and internet companies. Thus, providing SMART METER service using satellite is ideal for rural or difficult locations, and it is also possible to application in urban areas as well. There is a growing trend towards the use of TCP/IP technology as a common communication platform for Smart Meter applications, so that utilities can deploy multiple communication systems, while using IP technology as a common management platform.

1) Characteristics:

Some characteristics of fixed network scenario are:

a) Higher bandwidth availability for forward link and limited bandwidth in the return link

b) Multiple security domains

c) Wider coverage area comparing to mobile scenario

2) Security features:

Some security features of fixed network scenario are:

a) multiple encryption regions due to the multiple administrative domains

b) moderate/frequent data key updates might be necessary due to higher data rate in the forward link

c) For multicast services, slow/static group membership due to the nature of fixed network terminals

d) Centralised key/policy management architecture is the preferred solution.

e) Access control is one of the major concerns of SMART METER application. It is to ensure only devices authorised by the customer and energy supplier are allowed to interact with metering system.

f) How to manage and use the data key is essential. For the broadband services, the main types of communication that are supported by fixed network are voice, data transfer, video/images and web browsing. It might not be necessary to use data key to secure all of the traffic (such as large volumes of multimedia traffic). How to define the security level of different traffic becomes a challenge. For the SMART METER application, each meter, especially the sensitive user ID and billing related information, must be reliably and securely transferred to the central location.

C. DTN Scenario: Deep Space

Space exploration started in early sixties and since then the interest towards deep-space communication continuously increased especially from the scientific point view, thus paving the way for the Moon human exploration and then the Mars missions. More recently, the current advances and trends in technology have pushed the space agencies to a new and more futuristic concept of space exploration: the Solar System Internet. In fact, it consists in the deployment of a real Internet over the space, able to connect Earth centers to remote sites, located in possibly different places of the Solar System, such as Mars, Saturn, and Mercury. Consequently, it is immediate to think of a complex deep-space network (Figure 4), where data transaction and routing operations are performed seamlessly and autonomically, thus reducing the manual intervention to the least. The human assistance would be still needed to provide recovery to emergency situations that the implemented fault resilience model could not handle. Besides the attracting perspective, this future scenario may offer benefits in terms of scientific studies and possible revenues for the aerospace industries.

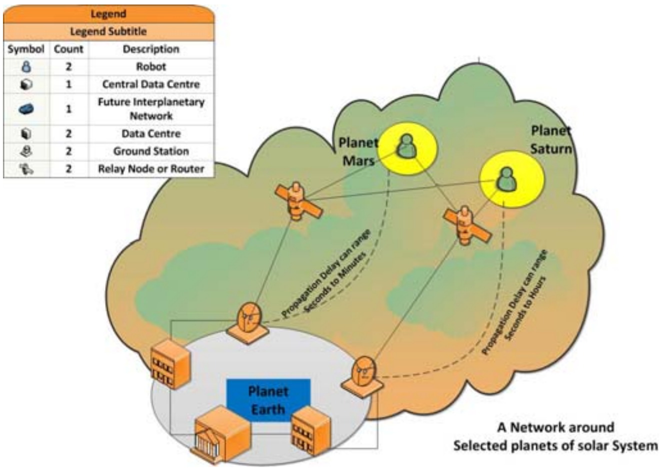

Legend
Legend Subtitle
Symbol
Count
Description
2
Robot
1
Central Data Centre
1
Future Interplanetary Network
2
Data Centre
2
Ground Station
2
Relay Node or Router
Planet Mars
Planet Saturn
Propagation Delay can range Seconds to Minutes
Propagation Delay can range Seconds to Hours
Planet Earth
A Network around Selected planets of solar System

Figure 4. DTN scenario: A Diagram showing a future network along planets of solar systems

1) Characteristics:

Some characteristics of DTN scenario are:

a) Extremely limited bandwidth availability

b) Limited number of security domains.

c) Frequent disconnection/disruptions

d) Very large propagation delays. Depending on the specific addressed space mission, the propagation delay can range from a few seconds (e.g., Earth-Moon) to several minutes (e.g., Earth-Mars), to even hours (e.g., Earth-Saturn, Earth-Pluto).

e) Scarce and highly asymmetric link data rate. Because of the reduced spacecrafts' size, the deployed antenna can be only of reduced dimensions, thus implying small data rate available. In addition, most of part of data traffic flows though the downlink (e.g., measurement, image transfer), whereas the uplink is principally used for transmitting telecommand messages. As a result, strong asymmetry between data rates available on downlink and uplink respectively is experienced, being as high as 10000 to 1.

f) Limited storage availability. The limited dimensions of the space crafts pose additional constraints on the on-board storage, which plays some role for routing and buffering.

g) Degraded link quality. The long distances determine high free-space-loss to which also weather fading may add, occurring in case of Ka band transmission. Besides, in case of optical laser technology, additional quality impairments may take place, resulting in non negligible BER or PER.

h) Intermittent visibility between Earth and other remote planets, because of the relative movement around the Sun, resulting in tight transmission schedule to take advantage of the available resources. Finally, this leads to an overall reduced throughput measure, if compared to the total mission time. However, by using the relay nodes or routers in the space, increased data rate and more communication opportunities can be achieved by using DTN store and forward mechanism

2) Security features:

Some security features of DTN scenario are:

a) Limited number of encryption regions, due to the nature of space application

b) Slow data key updates

c) For multicast services, slow changing/static group membership

d) Distributed key/policy management architecture is the preferred solution, due to the sparse nature of space communications..

VI. CONCLUSION

While the advantages of multicasting services over satellite networks are clear, security as one of the obstacles poses great challenges in terms of scalable key management and adaptable policy provisioning. Innovative security architecture is proposed in this paper to address the security challenges, with a particular focus on key management and security policy. The major issues on multicast key management/security policy are discussed. A brief literature review is provided and existing problems are highlighted. Also, three scenarios are defined for future implementation: mobile network scenario for the application such as mobile broadband, fixed network scenario for the application such as SMART METER/GRID and DTN scenario for the application of Deep Space. The characteristic of each scenario is analyzed and security requirements are also drawn.

Based on the security architecture, protocols between key managers, policy server and data distributor need to be defined in the future. Group Secure Association Key Management Protocol (GSAKMP) in [14] provides secure communications between group owner, key mangers, senders and receivers. Either GSAKMP-type protocol will be used to establish secure communications between data distributors the other entities or a new protocol will be developed, depending on the architecture requirements. If a new protocol is required, the proposed protocol will be analyzed and verified by model-checking or theorem-proving techniques.

REFERENCES

[1] I. Brown, J. Crowcroft, et al., "Internet Multicast Tomorrow", Internet Protocol Journal, 5(4), 2002.

[2] C. Diot, B.N. Levine, B. etal., "Deployment Issues for the IP Multicast Service and Architecture", IEEE Network, vol. 14, pp. 10-20, Jan/Feb, 2000.

[3] Y. Challal, etal., "Adaptive clustering for Scalable Key Management in Dynamic Group Communications", International Journal of Security and Networks, SSN 1747-8413, 2007.

[4] S. Rafaeli and D. Hutchison, "A Survey of Key Management for Secure Group Communication", ACM Computing Surveys, Vol. 35, No. 3, pp. 309–329, September 2003.

[5] R. Wittmann, M. Zitterbart, "Multicast communication: Protocols and applications", Morgan Kaufmann, ISBN 1-55860-645-9, 2001.

[6] K. Koyama and K. Ohta, "Identity-based conference key distribution systems," in Advances in Cryptology - CRYPTO'87 (LNCS No. 293), pp. 175--184, Springer-Verlag, 1987.

[7] A. Ballardie, "Scalable Multicast Key Distribution," RFC 1949, IETF, 1996.

[8] M. Steiner, et al., "Diffie-Hellman key distribution extended to group communications," in Proceedings of the 3rd ACM Conference on Computer and Communications Security, ACM, March 1996.

[9] H. Cruickshank, etal., "Securing multicast in DVB-RCS satellite systems". IEEE Wireless Communications, Special Issue on Key Technologies and Applications. October 2005

[10] D. Ng, H. Cruickshank, Z. Sun and M.P. Howarth, "Dynamic Balanced Key Tree Management for Secure Multicast Communications". IEEE Transactions on Computers. 2007

[11] Y. Zhang, A multilayer IP security protocol for TCP performance enhancement in wireless networks. IEEE Journal on Selected Areas in communications,vol. 22, no. 4, May 2004.

[12] T. Hardjono, et al., The Multicast Group Security Architecture. RFC 3740. 2004

[13] H. Harney and E. Harder, "Multicast Security Management Protocol (MSMP) Requirements and Policy", IETF, March 1999. draft-harney-sparta-msmp-sec-00.txt.

[14] H. Harney, et al., GSAKMP: Group Secure Association Key Management Protocol. RFC 4535. 2006

[15] T.Christian, M. Riguidel, "Distributed trust infrastructure and trust-security articulation: Application to heterogeneous networks". IEEE, Proceedings – 20th International Conference on Advanced Information Networking and Applications, 2006, p 33-38

[16] H. Harney, et al., "Principles of Policy in Secure Groups", Proceedings of Network and Distributed Systems Security 2001 Internet Society. 2001.

[17] M. Baugher, et al., Multicast Security Group Key Management Architecture. RFC 4046. 2005

[18] A. Colegrove, H. Harney, "Group Security Policy Token v1", RFC 4534, IETF, June 2006.

Energy-Efficient Computation of L1 and L2 Norms on a FPGA SIMD Accelerator, with Applications to Visual Search

Calin Bira*, Radu Hobincu*, Lucian Petrica*, Valeriu Codreanu†, and Sorin Cotofana‡
*Politehnica University of Bucharest
{calin.bira, radu.hobincu, lucian.petrica}@arh.pub.ro
†Eindhoven University of Technology
v.codreanu@tue.nl
‡Delft University of Technology
s.d.cotofana@ewi.tudelft.nl

Abstract—This paper presents a novel accelerator architecture which is SIMD in nature and fully programmable. It provides support in an energy effective manner to a wide range of vector computations, including scalar products and similarity metrics like sum of absolute differences and sum of squared differences. We have evaluated an implementation of the proposed architecture on the Xilinx Zynq-7000 EPP featuring the ARM Cortex-A9 processor, running a SIFT descriptor matching benchmark. Our results indicate that the processor can offload the most intensive computational kernels of the benchmark to the accelerator, thus delivering 4-6x better matching throughput than the ARM processor alone. Moreover, the execution of the SIFT matching benchmark on the accelerated platform consumes 3x less energy than on the ARM Cortex-A9, at a similar power consumption. Our results also suggest that the accelerated ARM system is 40% more energy effective than Intel Core i7 2600K and Nvidia GTX680 when executing the SIFT matching benchmark.

I. Introduction

Object recognition and classification are currently some of the hot topics in computer vision, with applications in image matching [11], robotics [16], and panorama stitching [4]. When matching large databases against each-other, matching speed is the most important performance metric, but power and energy efficiency plays a major role in the economy of the entire process. For robotics and mobile devices in general, energy efficiency is the most important metric since it relates directly to battery drain. Previous work has yet to demonstrate a solution to the image matching problem which is high-speed, low-power, and low-energy. Our research aims to prove that these goals are attainable without sacrificing programmability.

In this paper we propose an energy-efficient solution to the matching problem based on a Single Instruction Multiple Data (SIMD) accelerator architecture. The proposed architecture is well suited for execution of multiply-accumulate operations and for selective execution on large data vectors. It consists of an array of efficient processing elements which are fed instructions and data by the host processor, through the use of Direct Memory Access (DMA) and several FIFO interfaces. The proposed architecture was implemented and evaluated on the Xilinx Zynq-7000 EPP [15] running a SIFT descriptor matching benchmark on a standard image dataset. Our results indicate that the ARM host processor included within the Zynq-7000 can efficiently offload computationally expensive kernels to the accelerator, resulting in 4-6x better matching throughput than when executing alone. Also, the SIFT matching benchmark execution consumes 3x less energy on the proposed platform than on the ARM Cortex-A9 alone, at similar power consumption. Comparisons with desktop parts suggest that the accelerated ARM system is 40% more energy effective than a high-end desktop CPU-GPU system.

This paper is organized as follows. Section II presents details on image matching metrics and the computational requirements involved. Section III introduces the proposed architecture and programming model. In Section IV we present the implementation of the proposed architecture on an off-the-shelf programmable chip. In Section V we introduce the experimental results, with Section VI presenting some concluding remarks.

II. Image Matching

The object recognition process works in several steps. First, the images are split in two sets: the *query* and *search* images. The *search* images are the ones in which the objects are to be detected, while the *query* image contains the objects that we wish to find in the *search* set. The object recognition system does not work on the images themselves, but rather on a set of local features representing interesting characteristics of objects present in the image [17][13]. Therefore, the next step is to extract these local features, called *keypoints*, using algorithms such as Scale Invariant Feature Transform (SIFT) [12]. The third and final step is to find matching keypoints, which are identical or very similar in both the *query* image and at least one *search* image.

The keypoint matching task relies on finding the nearest neighbour of a given query keypoint, in the database of search keypoints, according to a certain distance metric. Formally, the nearest neighbour search problem is finding the element $NN(X)$ in a finite set Q included in a D-dimensional space minimizing the distance to the input vector X. This is described in Equation (1), where $argmin\ d(X,Y)$ is the tuple (X,Y) which minimizes function d.

$$NN(X) = \underset{Y \in Q}{\operatorname{argmin}}\ d(X,Y) \quad ; X \in \mathbb{R}^D, Y \in Q \subset \mathbb{R}^D \quad (1)$$

The distance metric $d(X,Y)$ is the L_p norm computed as in Equation (2). The most usual norma are SAD, i.e, L_1 and SSD, i.e, L_2. For a given matching task with Q keypoints in the query set and S keypoints in the search set, we need a total of $Q * S$ SSD or SAD operations to find the nearest neighbour from the search set for all query keypoints. A match is declared between a query keypoint and its nearest neighbour from the search set if the distance between them is below a given application dependent threshold.

$$L_p = (\sum_{i=1}^{D} |X_i - Y_i|^p)^{\frac{1}{p}} \quad (2)$$

For SIFT and similar algorithms, keypoints are described by large, typically 64 to 128 elements, vectors of parameters. The large size of these vectors increases the computational effort and memory required for the nearest-neighbor calculation. The SSD metric is more precise but makes intensive use of multiplication, which is either slow or requires more hardware resources, while the SAD metric can be implemented with only an adder, but requires a conditional execution based on which of the operands is larger. Both metrics require the accumulation of entire data vectors, which can be implemented on a scalar processor in a time proportional to the vector size, therefore slow for SIFT descriptors.

Research projects, e.g, demoASIFT [14], perform keypoint matching on general-purpose CPUs and support OpenMP [6] and vectorization, where available, to increase performance. There have been several proposals for nearest neighbour search using specialized hardware, such as GPUs. Garcia *et al.* and other research groups have implemented GPU nearest-neighbour search and have achieved as much as 120x speedup when comparing to a similar algorithm running on the CPU [10][5]. However, GPU implementations are extremely power-hungry, even though the high attainable matching speeds make for good energy efficiency.

With regard to specialized accelerators for similarity matching, Wong *et al.* demonstrated FPGA-based SAD implementations [20]. Their analysis indicates that an adder tree is the most efficient implementation, but while the achieved throughput is impressive, the accelerator is not programmable, and therefore is only useful for SAD computation. Flatt *et al.* proposed and evaluated a host-coprocessor scheme based on a RISC CPU coupled to an application-specific FPGA circuit and demonstrated 70x speedup when compared to RISC-only execution for the SSD metric [9]. Their approach also makes use of a fully pipelined adder tree architecture, and while configurable at block level, it is still not programmable.

The architectural solution proposed in the following section aims to increase the computational performance of L_1/L_2 matching algorithms on embedded systems, hence with minimal energy consumption, and also without sacrificing programmability.

III. Accelerator Architecture

This section presents the proposed system from a hardware and programming model perspective. The proposed SIMD architecture as well as its FPGA implementation are introduced below. The software section describes the programming environment and the corresponding software stack.

A. Hardware Architecture

The proposed system follows the host-accelerator paradigm . The accelerator is connected to the same bus as the CPU, the memory, and the Direct Memory Access (DMA) engine, and it is mapped to the processor address space. The processor can transfer data to the accelerator from its internal registers and memory, or it can instruct the DMA engine to transfer data directly from the main memory to the accelerator.

The SIMD accelerator, presented in Figure 1, follows the basic principles of the Connex multimedia processor [18]. The computation core consists of N Processing Elements (PEs), which are simple processors containing several internal registers, an Arithmetic Logical Unit (ALU), and instruction decoding logic. The instruction set follows the RISC principle, whereby all instructions execute in one clock cycle, operate on two registers and write the result to a third register. In order to speed up the computation, each PE benefits from a Local Storage (LS) similar to the shared memory employed by modern GPUs. The transfer between this shared memory and the system bus is done through an Input/Output (IO) network. The PEs are all fed the same instruction at any given time, which minimizes the control path overhead present in traditional architectures. Also, specific PEs can be marked as inactive based on arithmetic flags such as Carry, Less, and Equal. This permits the SIMD engine to selectively execute instructions, which is useful for computing certain types of matching metrics, e.g., SAD. Apart from the number-crunching PEs, the SIMD accelerator is composed out of three separate networks:

- The Input-Output network for controlling I/O data transfers between the main memory and the Local Storage.
- The Distribution network for dispatching the to be executed instructions to the PEs, in the form of a fully pipelined logarithmic tree.
- The Reduction network for gathering the result of global operations like sum reduction, in the form of a fully pipelined adder tree.

Mechanisms, e.g., instruction fetching, branching, and scalar computation, implemented in hardware in the Connex processor, have been moved into software, in order to minimize area and power overhead associated with control functions. A simple loop sequencer is retained to execute loops with a constant number of iterations.

The structure of the SIMD accelerator closely matches the requirements of nearest-neighbour search with the L_1 and L_2

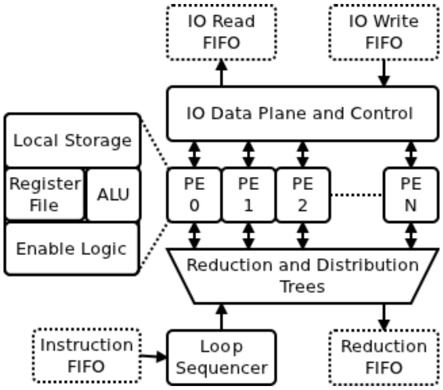


Fig. 1. SIMD Engine Architecture.

norms, which relies on two stages: an element-by-element operation, followed by a reduction operation. The first stage involves performing the same instruction on multiple chunks of data, and hence an SIMD architecture can efficiently do the computation on this level. To speedup the SSD evaluation, each PE includes a hardware multiplier, while the conditional execution feature is targeted at efficient SAD computation. Multiplication is done in two separate instructions: the first launches the multiplication while the second moves the result into the destination register, thus keeping with the RISC principle.

The second computation stage involves the summation of all results computed during the first stage. In our architecture this is implemented with a fully pipelined adder tree, which was found in [20] to be the most efficient way to implement sum reduction for SAD, and was also utilized for SSD in [9].

B. Programming Model

The software architecture and programming model were developed for easy integration with the user code. Instruction fetch as well as most control instructions handled by the host processor, using our accelerator-specific library called OPINCAA (Opcode Injection and Control for Accelerator Architectures).

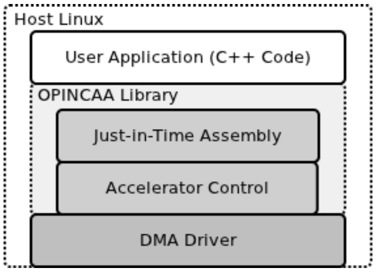


Fig. 2. Accelerator Software Stack.

The overall software architecture is depicted in Figure 2. Accelerator code, on the highest level of the hierarchy, is written in C++ in a specific syntax and can be mixed with ordinary C++ code in order to implement loops and branches. For this purpose, OPINCAA provides a vector data type and operators specific to it, including reduction, cell selection and arithmetic operations. The vectors used in OPINCAA always have the same size as the underlying accelerator implementation.

Kernels of accelerator code are compiled on-demand in a similar fashion to just-in-time (JIT) compilation for Java code. OPINCAA provides the JIT infrastructure as well as other accelerator control functions. Data-dependent accelerator loops are unrolled in software, while loops without data dependencies, and constant number of iterations, are preserved and executed in hardware by the loop sequencer. Compiled kernels are indexed and stored until the user explicitly requests their execution with a call to *executeKernel(index)*, which is a function of the accelerator control driver. In OPINCAA, instruction dispatch is a write to a device file, as are IO writes. Similarly, reductions and IO data are read through file interfaces. The control driver also exposes the functions *readReduction()*, *ioRead()*, and *ioWrite()*, which ensure the correct usage of the file interfaces. These functions abstract the operation of the underlying DMA driver from the user. The DMA driver is implementation-specific, as is the accelerator hardware.

Algorithm 1 SIMD code example.

```
$for(int\ i = 0;\ i < 10;\ i++)\{$
    $R[1] = LS[i];$
    $R[2] = LS[i + 10];$
    $R[3] = LS[R[1]];$
    $R[0] = (R[1] - R[2]);$
    $WHERE\_EQUAL($
        $R[3] = R[1] + R[2];$
    $)$
    $REPEAT(5)$
        $R[3] = R[3] + R[2];$
        $REDUCE(R[3]);$
    $END\_REPEAT$
$\}$
```

Algorithm 1 represents a code snippet to illustrate the syntax used to program the accelerator. The first two lines load the contents of the Local Storage at address i into $R[1]$ and address $i + 10$ into $R[2]$. The third line loads the value at an address stored in a register. The following line computes the difference between $R[1]$ and $R[2]$, places the result in $R[0]$ and sets the appropriate flags. The Carry flag indicates if the sum of $R[1]$ and $R[2]$ overflows, Less indicates if $R[1]$ is less than $R[2]$ and Equal indicates if $R[1]$ and $R[2]$ are equal. The $WHERE$ construct deselects the processing elements where the indicated flag is not set. Execution continues only in selected PEs until an implicit instruction is encountered to re-enable all PEs. In the case of Algorithm 1, the PEs where $R[1]$ and $R[2]$ are equal load their sum into $R[3]$, while all others keep the initial value loaded from the Local Storage. Finally, several summation and reduction operations are launched on register $R[3]$ through the use of the $REPEAT$ statement, which signals a loop executed in hardware by the loop sequencer. The outer loop, in plain C++ syntax, is unrolled during the JIT assembly, before the program is streamed to the accelerator. Also, any non-vector variable which is used in vector code is replaced by its value during the JIT assembly, and treated as a constant thereafter.

IV. Implementation

This section describes an implementation of the proposed architecture on the Xilinx Zynq-7000 extensible processing platform [15]. The platform consists of two parts, namely the Processing System (PS) and the Programmable Logic (PL). The PS includes a dual-core ARM Cortex-A9 processor running at 667Mhz with NEON instruction support containing also several I/O peripherals. The PS includes a DDR SDRAM memory controller and can boot independently of the programmable logic. The PL consists of a full-fledged Artix-7 FPGA fabric. The PS is linked to the PL through an AMBA AXI bus, and hence the FPGA fabric can accommodate digital circuits that accelerate the computation performed by the ARM cores. For high-speed data transfer between the PS and PL, the Zynq architecture includes a Direct Memory Access (DMA) engine which can be programmed by the ARM processor.

An instance of the architecture was implemented on the Zynq, with the following parameter values: 128 Processing Elements, 16-bit operands, 32 registers, and 2KB Local Storage. This system instance came out of the need to perform fast and efficient nearest-neighbour computation in high-dimensional spaces for SIFT primarily, since it is considered the best-performing matching algorithm. SIMD sizes of 128 processing elements effectively permit calculating the distance between all 128 elements of two SIFT descriptors in one pass. The 16-bit operand dimension was chosen because it has already been used in [14]. Moreover the work in [19] indicates that using a shorter integer representation (short int in our case) instead of full integers or float operands does not result in significant loss of matching accuracy.

The size of the Local Storage was chosen as the size of a Block RAM resource of the Zynq FPGA, while the number of registers allows for an efficient register file in Distributed RAM. Hardware multipliers were implemented in DSP48E1 slices. The resulting accelerator design occupies 90% of the Zynq FPGA and can be clocked at 125Mhz. In our experiment, a 100Mhz frequency was utilized in order to match the AXI bus frequency.

Figure 3 illustrates how the Zynq-7000 is utilized within our approach. The PS connects to the PL through a Xillybus interface core [1], which makes use of the DMA engine to transfer data from the main memory to several FPGA FIFOs. The SIMD accelerator connects to these FIFOs, consuming and producing data from and to the FIFOs as instructed by the PS. On the Zynq, OPINCAA makes use of the Xillybus DMA driver, which abstracts transfers over AXI and exposes the required file interfaces.

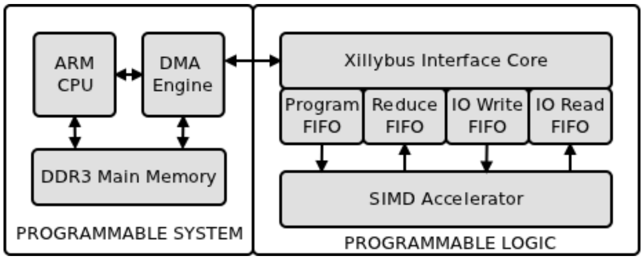


Fig. 3. Zynq-7000 Accelerator.

V. Experimental Results

In order to compare our approach with other matching hardware, we integrated our keypoint matching functions within the demoASIFT project [14]. The evaluation results are obtained from a Zedboard [2] development board housing the Xilinx Zynq-7000 EPP, as presented in Section IV. The benchmark was compiled with gcc version 4.6, which is provided with the Xillinux operating system.

To accurately benchmark our proposal, we made us of images from the standard dataset proposed by Mikolajczyk *et al.* to evaluate feature extractors [13]. The feature extraction was done on the Zynq PS, while the matching was executed either on the PS or on the SIMD accelerator configured on the Zynq PL. We measure the execution time, as well as the energy consumption required to perform image matching for two systems:(i) the baseline ARM system and (ii) the SIMD accelerated system where computation is split between the ARM and the accelerator. Energy consumption for Intel Core i7 2600K quad-core desktop CPU, NVidia GTX680 GPU, and NVidia 8800 Ultra GPU are also presented for the purpose of comparison.

The benchmark code in [14] allows vectorization with Intel SSE [8] by default. We have modified the benchmark in order to make automatic vectorization possible on ARM with NEON [7] instructions, thereby extracting the maximum speed out of the Zynq PS. OpenMP was also used to split the matching workload on all the available processor cores.

A. SIMD Matching Algorithms

By using the previously described SIMD accelerator, we can code the SIFT matching application using Algorithm 2 for SSD or Algorithm 3 for SAD. As discussed in Section II, we need to do $Q*S$ SSD/SAD operations to match a query set of Q keypoints to a search set of S keypoints. For efficient use of the LS and registers, the search and query keypoint sets are broken up into tiles of 308 and 364 keypoints, respectively, in order for one query tile and two search tiles to fit inside the LS simultaneously. One tile from the query set is loaded into the LS and tiles from the search set are sequentially loaded and matched against it. Processing is done on sub-tiles of 28 keypoints from the search set, which allows the register file to be fully used. While processing occurs on one search tile, a second tile is loaded into the LS, thus masking most of the IO time behind the computation.

Algorithm 2 SSD Matching Kernel.

```
Require: keypoints to be transferred to LS
  $BEGIN\_KERNEL();$
    $R29 = 1;$
    $for(int\ i = 0;\ i < 11;\ i++)\{$
      $for(int\ j = 0;\ j < 28;\ j++)\{$
        $R[j] = LS[364 + ls\_off * (28 * 11) + i * 28 + j];$
      $\}$
      $R30 = 0;$
      $REPEAT(364)$
        $R[28] = LS[R30];$
        $R30 = R30 + R29;$
        $for(int\ j = 0;\ j < 28;\ j++)\{$
          $R31 = R[28] - R[j];$
          $R31 = R31 * R31;$
          $REDUCE(R31);$
        $\}$
      $END\_REPEAT$
    $\}$
  $END\_KERNEL();$
```

The actual SSD computation is contained within the innermost loop of Algorithm 2. A search keypoint is loaded from the LS, is subtracted from the currently selected query keypoint, and the difference is squared before being launched in the reduction network, which does the summation. Algorithm 3 presents the innermost loop of the SAD accelerator kernel, which uses conditional execution. In this case R29 is zero and is used to test whether the result of the difference needs to be negated. This code hides some of the instructions actually in use, which re-enable the PEs after the conditional execution. The extra instructions make SAD computation slower than SSD on our accelerator, despite the fact that multiplication takes two cycles to complete.

Algorithm 3 SAD Computation.

```
  $for(int\ j = 0;\ j < 28;\ j++)\{$
    $R30 = R[28] - R[j];$
    $R31 = R30 < R29;$
    $WHERE\_LT($
      $R30 = R[j] - R[28];$
    $)$
    $REDUCE(R31);$
  $\}$
```

B. Performance

Table I presents the performance results in terms of millions of keypoint matches per second (MM/s), where a keypoint match is an SAD or SSD operation on a pair of SIFT keypoints. The SIMD accelerated implementation provides a speedup of 3.94 and 6.35 for SAD and SSD, respectively, over the baseline implementation. We note inhere that during the baseline ARM-only execution, both Cortex cores are fully utilized, while when executing with the use of the SIMD accelerator, a single core is utilized at around 65%. Also, the NEON resources on the Cortex cores are idle during the accelerator enhanced execution.

TABLE I
SSD AND SAD MATCHING.

Platform	ARM Cortex A9	SIMD Accelerator
Frequency [MHz]	667	100
SSD Rate [MM/s]	2.11	13.40
SSD Speedup	1	6.35
SAD Rate [MM/s]	2.34	9.22
SAD Speedup	1	3.94

The SIMD accelerator performs significantly better on the SSD metric, as it does not require conditional execution. Setting up and disabling conditional execution requires extra clock cycles to be used when computing the SAD metric. On the ARM, however, the nature of NEON vector instructions causes SSD execution to be about 10% slower than that of SAD, which is the opposite of what we observe on the accelerator. This happens because of the built-in support for the *VABA* (Vector Absolute Difference and Accumulate) instruction.

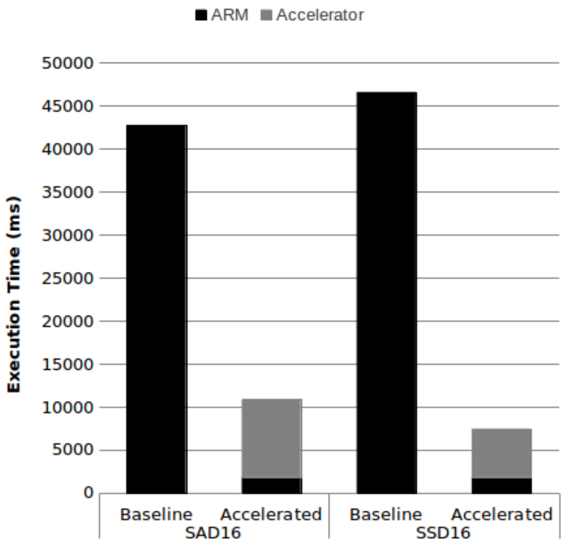


Fig. 4. Profiling of Execution Time.

Figure 4 presents an execution time break-down corresponding to the baseline and accelerated matching. We can observe that even during accelerated matching, a significant proportion of the execution time is occupied by tasks executed on the ARM processor. These are mainly tasks related to decisions on the rejection of matching pairs of keypoints in some circumstances. Further speedup could be attained by moving these decision processes to a separate thread and performing them while matching data are received from the accelerator.

C. Energy

The dual-core ARM CPU consumes, according to Zedboard documentation, a maximum of 1.25 Watts, yielding on average

TABLE II
ENERGY CONSUMPTION PER 100 MMATCHES.

Platform	TDP[W]	SAD energy [J]	SSD energy [J]
Core i7 2600K	95	83.77	76.98
NVidia GTX680	195	24.23	24.37
NVidia 8800 Ultra	175	–	286.88
ARM Cortex A9	1.25	53.41	59.24
SIMD accelerator	1.2	13.01	8.95

2.23 MM/s. The SIMD accelerator consumes 600 mW according to Xilinx power estimation tools. To this we must add the power consumed by the programmable system to control the accelerator. In total, the accelerator consumes approximately 1.2 Watts. To evaluate the implications of our proposal in terms of energy consumption, we calculate and present in Table II the energy per 100 MM for our scheme and equivalent state of the art implementations. This is by no means a comprehensive energy evaluation, but it gives an estimate of the energy-efficiency of the proposed solution.

The Intel CPU results are measured from demoASIFT, with automatic vectorization on SSE and OpenMP used for parallelization. The GPU results for the NVidia 8800 Ultra on SSD matching were extracted from previous work [5]. We have used the cv::gpu::BruteForceMatcher class from OpenCV 2.4 [3] to measure the performance of the GTX680. It must be noted that for the GPU cases, only the device Thermal Design Power (TDP) was taken into account, while the power required by the host for control functions was ignored. In this case we feel that the use of TDP instead of measured power is justified since the chips are fully utilized, including vector resources for the CPU parts, while running the matching benchmarks. Table II suggests that the most energy efficient platform is the accelerated Zynq system, followed by the GTX680, and the baseline Zynq system. The GTX680 performs well because it can exploit the large amount of parallelism in the matching application and delivers 800 MM/s on both metrics, thereby compensating for its high power consumption. The baseline Zynq system is competitive with regard to energy because it consumes very little power. The 8800 Ultra GPU is an older generation chip and its performance is much less than the GTX680, resulting in higher energy consumption. The proposed accelerated system is at least two times more energy efficient than the other platforms.

VI. CONCLUSIONS

We have presented a hardware SIMD accelerator architecture specifically tailored for similarity matching in computer vision algorithms. The accelerator is designed to work in conjunction with an embedded processor and enable high matching throughput for mobile applications energy-constrained applications like robotics. We have implemented this architecture in the Zynq-7000 system-on-chip on the Zedboard development platform, using a Xillybus core for data transfer between the ARM processor and the accelerator. Evaluation has revealed that the SIMD accelerator is able to achieve 4-6x better SIFT descriptor matching throughput than a Cortex A9 processor, despite the FPGA implementation and 100MHz operating frequency. This performance is delivered at roughly 3x less energy consumption and similar power consumption. The accelerated system is 40% more energy effective even than Intel Core i7 2600K and Nvidia GTX680 when executing the SIFT matching benchmark.

ACKNOWLEDGEMENT

Part of this work was carried out with funding and support from POSDRU/159/1.5/132397 ExcelDoc program.

REFERENCES

[1] Xillybus. ”http://xillybus.com”.

[2] Zedboard. ”http://www.zedboard.org/”.

[3] Gary Bradski. The OpenCV library. *Doctor Dobbs Journal*, 25(11):120–126, 2000.

[4] M. Brown and D.G. Lowe. Recognising panoramas. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 2, page 5, 2003.

[5] A. Chariot and R. Keriven. GPU-boosted online image matching. In *19th International Conference on Pattern Recognition*, pages 1–4, 2008.

[6] Leonardo Dagum and Ramesh Menon. OpenMP: an industry standard API for shared-memory programming. *Computational Science & Engineering, IEEE*, 5(1):46–55, 1998.

[7] Pierre Esterie, Mathias Gaunard, Joel Falcou, et al. Exploiting multimedia extensions in C++: A portable approach. *Computing in Science & Engineering*, 14(5):72–77, 2012.

[8] Nadeem Firasta, Mark Buxton, Paula Jinbo, Kaveh Nasri, and Shihjong Kuo. Intel AVX: New frontiers in performance improvements and energy efficiency. *Intel White paper*, 2008.

[9] Holger Flatt, Sebastian Hesselbarth, Sebastian Flügel, and Peter Pirsch. A modular coprocessor architecture for embedded real-time image and video signal processing. *Embedded Computer Systems: Architectures, Modeling, and Simulation*, pages 241–250, 2007.

[10] Vincent Garcia, Eric Debreuve, and Michel Barlaud. Fast k nearest neighbor search using GPU. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW’08. IEEE Computer Society Conference on*, pages 1–6. IEEE, 2008.

[11] MC Kus, M. Gokmen, and S. Etaner-Uyar. Traffic sign recognition using Scale Invariant Feature Transform and color classification. In *23rd International Symposium on Computer and Information Sciences, ISCIS’08*, pages 1–6. IEEE, 2008.

[12] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[13] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and L Van Gool. A comparison of affine region detectors. *International journal of computer vision*, 65(1):43–72, 2005.

[14] J.M. Morel and G. Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.

[15] Mike Santarini. Zynq-7000 EPP sets stage for new era of innovations. *Xcell journal*, 75:8–13, 2011.

[16] S. Se, D. Lowe, and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *IEEE International Conference on Robotics and Automation, Proceedings 2001 ICRA*, volume 2, pages 2051–2058. IEEE, 2001.

[17] Linda Shapiro and George C Stockman. *Computer Vision. 2001*. Prentice Hall, 2001.

[18] Gheorghe Stefan. The CA1024: A massively parallel processor for cost-effective HDTV. In *Spring Processor Forum: Power-Efficient Design*, pages 15–17, 2006.

[19] Tinne Tuytelaars and Cordelia Schmid. Vector quantizing feature space with a regular lattice. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[20] Stephan Wong, Bastiaan Stougie, and Sorin Cotofana. Alternatives in fpga-based sad implementations. In *Proceedings of the IEEE International Conference on Field-Programmable Technology*, pages 449–452, 2002.

An Engine Oil Replacement Timeline

José Neves, Filipa Ferraz, Henrique Vicente and Paulo Novais

***Abstract*—**Engine oil conditions are crucial either for the proper engine function and to calculate the oil´s lifetime. Factors like rotation per minute, temperature, trip length, clarity of the oil is important to determine the level of contamination of the engine oil. Thus, the type of engine heating cycles, the power of the engine and the components of the oil used are also factors that determine whenever oil replacement should occur. Therefore, analyzing oil contaminations matters when it comes to replace it again. Consequently, manufactures and mechanics recommendations are not always the best, leading to consequences such as corrosion and sludge in the engine. So, this work focuses on engine oil parameters and environmental conditions to estimate the optimal oil replacement intervals, here given in terms of a replacement timeline, which may also contribute to a better environment that embodies all living and no living things on Earth.

***Keywords*—**Artificial Neuronal Networks, Energy and Environment, Engine Oil Replacements, Logic Programming.

I. Introduction

OIL chemistry and engine technology have evolved impressively in recent years, but nobody would ever know it from the behavioral changes of car owners, and its impact on the environment. Mostly driven by an outdated oil change commandment, they are needlessly dripping an ocean of contaminated waste oil.

On the one hand the majority of automakers today call for oil changes at either 10000 or 15000 kilometers, and the interval may go as high as 25000 kilometers in some cars. Yet, this wasteful cycle continues largely because the automotive service industry, while fully aware of the technological advances, continues to address the 5000-kilometre gospel as a way to keep the service bays busy. As a result, car owners are dumping their engine oil twice as often as their service manuals endorse.

On the other hand, the 5000-kilometre oil change is a fable that should be laid to cessation. Failing to heed the service interval in your owner's manual wastes oil and money, while compounding the environmental impact of illicit dumping. Part of the blame for this over-servicing lies in our insecurities about increasingly complicated engines that are all but inaccessible to the average driver.

Because busy car owners seldom read their owner's manuals, most have no idea of the actual oil change interval for their cars. Therefore, they blindly follow the windshield reminder sticker, whether it is an accurate indicator of the need for an oil change or not, i.e., the quality of oil has changed quite a bit, and the public in general is not aware of that [1]. Oil amendment interims are, mainly, due to:

- An improved robustness of today's oils;
- More and more automakers using synthetic oil;
- Tighter tolerances of modern engines; and
- The introduction of oil life monitoring systems, which tell the driver when an oil change is needed. However, existing systems do not address incomplete, contradictory or default information, which is their major drawback.

Indeed, the engine is like the heart of a car, i.e., it is an essential part of it. Thus, its components have to be in order and in accurately conditions, like human’s vessels. And one of those components it is the engine oil that allows the engine’s lubrication, so it can function properly.

Engine oil has some features that are relevant like viscosity high enough to maintain a greasing film, but, at the same time, low enough to let it flow throw the various parts of the engine at any condition. Therefore, the viscosity is seen as one of main parameters to check, especially when it is controlled by the temperatures variations at which the engine is subjected, by the oil composition, since the fact that it can be synthetic or not, petroleum-based or not, by atmospheric conditions, by the number of heating cycles of the engine, by the kind of use of the car, among others factors, that can lead to viscosity variations and oil contamination, making consequently repercussions in the engine and in its function [2], [3]. Still about oil contamination, there are issues like engine power, fuel type and trip average length that may according to its results, affect negatively the oil clarity, which means the presence of wastes in it [4].

Hence, it’s advisable an engine oil replacement between certain time intervals or mileage, based on real and specific parameters instead on a generic prediction that suppliers and manufactures make [5]. Really, as it was already stated above, they recommend a change of oil about every 10000 to 15000 kilometers [6]. Of course it depends on the factors referred to above, but these are the more common replacement intervals suggested by the oil producers and mechanics.

Therefore, the change-interval question: When to perform an oil change?

This work is funded by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within projects PEst-OE/EEI/UI0752/2014 and PEst-OE/QUI/UI0619/2012.

José Neves is with the CCTC/Department of Informatics, University of Minho, Braga, Portugal (corresponding author to provide phone: +351-934201337; fax: +351-253604471; e-mail: jneves@di.uminho.pt).

Filipa Ferraz is with the School of Engineering, University of Minho, Braga, Portugal (filipatferraz@gmail.com).

Henrique Vicente is with the Department of Chemistry & Évora Chemistry Centre, School of Science and Technology, University of Évora, Évora, Portugal (e-mail: hvicente@uevora.pt).

Paulo Novais is with the CCTC/Department of Informatics, University of Minho, Braga, Portugal (e-mail: pjon@di.uminho.pt).

When one is working on oil monitoring, the challenging question with which car owners are met relies on the similarity between the signs and symptoms among the quality of oil.

Typically, the earliest acts for the diagnosis of oil quality proceed from the detection of a specific outbreak. An oil analysis will tell you the condition of your oil, and it also can reveal any problems that your engine may be experiencing. Some sample tests can show traces of fuel and coolant in the engine oil, which are early signs of engine problems. These tests may close the door to mistake and doubt.

Above any kind of dispute, the main problem with the diagnosis of oil change comes from the large number of different states that may mimic its signs. Also this pinpointing is usually done over a large period of time, therefore generating a huge amount of data, which has to be treated and interpreted by the car owner(s) [7]–[10].

Facing with such a large amount of facts, even experienced experts have difficulties to make a precise diagnosis and distinguishing between this and other car malfunctions. With this article we make a start on the development of an unusual or original diagnosis assistance system for oil change. We will center on a logic programming based approach to knowledge representation and reasoning, complemented with a computational framework based on Artificial Neural Networks.

II. 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 [11]–[13], and Proof Theory [14], [15]. 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:

$$p \leftarrow p_1, \cdots, p_n, not\ q_1, \cdots, not\ q_m \tag{1}$$

$$?\,(p_1, \cdots, p_n, not\ q_1, \cdots, not\ q_m)\ \ (n, m \geq 0) \tag{2}$$

where *?* 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 $\neg$ [15]. Under this representation formalism, every program is associated with a set of abducibles [11], [13], 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.

Due to the growing need to offer user support in decision making processes some studies have been presented [16], [17], related to the qualitative models and qualitative reasoning in Database Theory and in Artificial Intelligence 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 [18], [19]. 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:

$$QoI_i = lim_{N\rightarrow\infty} \frac{1}{N} = 0 \qquad (N \gg 0) \tag{3}$$

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:

$$QoI_i = {}^{1}/_{Card} \tag{4}$$

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:

$$QoI_i = \frac{1}{C_1^{Card} + \cdots + C_{Card}^{Card}} \tag{5}$$

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.:

$$\sum_{1\leq k\leq n} w_i^k = 1, \forall_i \tag{6}$$

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 $x = (x_1, \cdots, x_n)$, defined in terms of the attributes of $predicate_i$, one may have:

$$V_i(x) = \sum_{1\leq k\leq n} w_i^k \times QoI_i\,(x)/n \tag{7}$$

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 *Manufacturer Specifications*, *Mechanic´s Observations*, *Oil Manufacturer Recommendations*, and *Engine Oil*.

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:

$$predicate_i\big((\,x_1, \cdots, x_n)\big) :: QoI \tag{8}$$

and evaluate the *Degree of Confidence* (*DoC*) given by $DoC = V_i(x_1, \cdots, x_n)/n,$ which denotes one‘s confidence on a particular term of the extension o$f\ predicate_i.$To be more general, let us suppose that the Universe of Discourse is described by the extension of the predicates:

$$a_1(\cdots), a_2(\cdots), \cdots, a_n(\cdots) \qquad (n \geq 0) \tag{9}$$

Therefore, for a given *scenario$_i$*, one may have (where $\perp$ denotes an argument value of the type unknown; the values of the others arguments stand for themselves):

$$\begin{cases} \neg a_1(x_1, y_1, z_1) \leftarrow not\ a_1(x_1, y_1, z_1) \\ \quad a_1(\perp,\ \ [10, 20], 15) :: 0.5 \\ \quad \underbrace{[5, 10][5, 30][10, 20]}_{\text{attribute`s domains for } x_1, y_1, z_1} \\ \\ \neg a_2(x_2, y_2, z_2) \leftarrow not\ a_2(x_2, y_2, z_2) \\ \quad a_2([45, 54], [10, 12], \perp) :: 0.65 \\ \quad \underbrace{[30, 60]\ \ [6, 14]\ [2000, 6000]}_{\text{attribute`s domains for } x_2, y_2, z_2} \\ \\ \qquad\qquad \vdots \end{cases}$$

$$\Big\Downarrow\ \textit{1st interaction: transition to continuous intervals}$$

$$\begin{cases} \neg a_1(x_1, y_1, z_1) \leftarrow not\ a_1(x_1, y_1, z_1) \\ \quad a_1([5, 10], [10, 20], [15, 15]) :: 0.5 \\ \quad \underbrace{[5, 10]\ \ [5, 30]\ \ \ [10, 20]}_{\text{attribute`s domains for } x_1, y_1, z_1} \\ \\ \neg a_2(x_2, y_2, z_2) \leftarrow not\ a_2(x_2, y_2, z_2) \\ \quad a_2([45, 54], [10, 12], [2000, 6000]) :: 0.65 \\ \quad \underbrace{[30, 60]\ \ [6, 14]\ \ [2000, 6000]}_{\text{attribute`s domains for } x_2, y_2, z_2} \\ \\ \qquad\qquad \vdots \end{cases}$$

$$\Big\Downarrow\ \textit{2nd interaction: normalization } \frac{Y - Y_{min}}{Y_{max} - Y_{min}}$$

$$\begin{cases} \neg a_1(x_1, y_1, z_1) \leftarrow not\ a_1(x_1, y_1, z_1) \\ \\ \quad a_1\left(\left[\frac{5-5}{10-5}, \frac{10-5}{10-5}\right], \left[\frac{10-5}{30-5}, \frac{20-5}{30-5}\right], \left[\frac{15-10}{20-10}, \frac{15-10}{20-10}\right]\right) \equiv \\ \\ \quad a_1([0, 1], [0.2, 0.6], [0.5, 0.5]) :: 0.5 \\ \quad \underbrace{[0, 1]\ \ \ [0, 1]\ \ \ \ [0, 1]}_{\text{attribute`s domains for } x_1, y_1, z_1} \\ \\ \neg a_2(x_2, y_2, z_2) \leftarrow not\ a_2(x_2, y_2, z_2) \\ \\ \quad a_2\left(\left[\frac{45-30}{60-30}, \frac{54-30}{60-30}\right], \left[\frac{10-6}{14-6}, \frac{12-6}{14-6}\right], \left[\frac{2000-2000}{6000-2000}, \frac{6000-2000}{6000-2000}\right]\right) \equiv \\ \\ \quad a_2([0.5, 0.8], [0.5, 0.75], [0, 1]) :: 0.65 \\ \quad \underbrace{[0, 1]\ \ \ \ \ [0, 1]\ \ \ \ [0, 1]}_{\text{attribute`s domains for } x_2, y_2, z_2} \\ \\ \qquad\qquad \vdots \end{cases}$$

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 te arguments intervals, once normalized.

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.

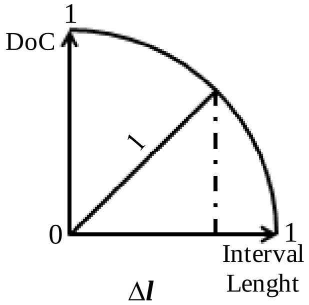


Fig. 1 Evaluation of *Degree of Confidence*

$$\begin{cases} \neg a_{1_{DoC}}(x_1, y_1, z_1) \leftarrow not\ a_{1_{DoC}}(x_1, y_1, z_1) \\ \quad a_{1_{DoC}}(0,\ \ \ 0.916,\ \ \ \ 1)\ \ ::\ \ 0.5 \\ \quad \underbrace{[0, 1]\ [0.2, 0.6]\ \ [0.5, 0.5]}_{\text{attribute`s values ranges for } x_1, y_1, z_1} \\ \quad \underbrace{[0, 1]\ \ \ [0, 1]\ \ \ \ \ \ [0, 1]}_{\text{attribute`s domains for } x_1, y_1, z_1} \\ \\ \neg a_{2_{DoC}}(x_2, y_2, z_2) \leftarrow not\ a_{2_{DoC}}(x_2, y_2, z_2) \\ \quad a_{2_{DoC}}(0.954,\ \ \ 0.968,\ \ \ \ 0)\ \ ::\ \ 0.6 \\ \quad \underbrace{[0.5, 0.8]\ [0.5, 0.75]\ [0, 1]}_{\text{attribute`s values ranges for } x_2, y_2, z_2} \\ \quad \underbrace{[0, 1]\ \ \ \ \ \ [0, 1]\ \ \ \ \ [0, 1]}_{\text{attribute`s domains for } x_2, y_2, z_2} \\ \\ \qquad\qquad \vdots \end{cases}$$

III. A Case Study

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 [20], 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 oil replacements intervals. Under this scenario some incomplete data is also available. For instance, in relation Mechanic´s Observation the value for Oil Replacement of model AXG11TRE is unknown, while in relation Oil Manufacturer Recommendations values for Mileage of model AXG11TRE range in the interval [10000, 15000].

Now, we may consider the extensions of the relations given in Fig. 2 to populate the extension of the *engine$_{oil}$* predicate, given in the form:

$$engine_{oil}: RPM, HP, Oil\ Capacity, Fuel, Trip\ Lenght, Oil\ Replacement, Oil\ Clarity, Mileage\ \rightarrow\ \{0,1\}$$

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 *engine_{oil}*, in the form:

$$\{$$

$$\neg engine_{oil}(RPM, HP, Cap_{acity}, F_{uel}, Trip_{lenght}, Repl_{acement}, Cla_{rity}, M_{ileage})$$

$$\leftarrow not\ engine_{oil}(RPM, HP, Cap_{acity}, F_{uel}, Trip_{lenght}, Repl_{acement}, Cla_{rity}, M_{ileage})$$

$$engine_{oil}(4600,\quad 94,\quad 6.5,\quad 0,\quad 1,\quad 10302,\quad 1,\quad 10000)\ ::\ 1$$

$$\underbrace{[4600,6500]\ [56,130]\ [3.5,8.5]\ [0,1]\ \ [0,1][10302,11686][0,1][8000,15000]}_{\text{attribute`s values ranges}}$$

$$engine_{oil}(5800,\quad 56,\quad 3.5,\quad 1,\quad 1,\quad \perp,\quad 0,\quad [10000,15000])\ ::\ 1$$

$$\underbrace{[4600,6500]\ [56,130]\ [3.5,8.5]\ [0,1]\ \ [0,1][10302,11686][0,1]\ \ [8000,15000]}_{\text{attribute`s values ranges}}$$

$$\}$$

In this program, the first clause denotes the closure of predicate *engine_{oil}*. The next clauses correspond to two terms taken from the extension of the *engine_{oil}* 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:

$$\{$$

$$\neg engine_{oil}(RPM, HP, Cap_{acity}, F_{uel}, Trip_{lenght}, Repl_{acement}, Cla_{rity}, M_{ileage})$$

$$\leftarrow not\ engine_{oil}(RPM, HP, Cap_{acity}, F_{uel}, Trip_{lenght}, Repl_{acement}, Cla_{rity}, M_{ileage})$$

$$engine_{oil}([0,0], [0.514,0.514], [0.6,0.6], [0,0], [1,1], [0,0], [1,1], [0.286,0.286])\quad ::\ 1$$

$$\underbrace{[0,1]\quad [0,1]\quad [0,1]\quad [0,1]\ [0,1]\ [0,1]\ [0,1]\quad [0,1]}_{\text{attribute`s domains}}$$

$$engine_{oil}([0.632,0.632], [0,0], [0,0], [1,1], [1,1], [0,1], [0,0][0.286,1])\quad ::\ 1$$

$$\underbrace{[0,1]\quad [0,1]\ [0,1]\ [0,1]\ [0,1]\ [0,1]\ [0,1]\quad [0,1]}_{\text{attribute`s domains}}$$

$$\}$$

The logic program referred to above, is now presented in the form:

$$\{$$

$$\neg engine_{oil_{DoC}}(RPM, HP, Cap_{acity}, F_{uel}, Trip_{lenght}, Repl_{acement}, Cla_{rity}, M_{ileage})$$

$$\leftarrow not\ engine_{oil_{DoC}}(RPM, HP, Cap_{acity}, F_{uel}, Trip_{lenght}, Repl_{acement}, Cla_{rity}, M_{ileage})$$

$$engine_{oil_{DoC}}(1,\quad 1,\quad 1,\quad 1,\quad 1,\quad 1,\quad 1,\quad 1)\quad ::\ 1$$

$$\underbrace{[0,0], [0.514,0.514], [0.6,0.6], [0,0], [1,1], [0,0], [1,1], [0.286,0.286]}_{\text{attribute`s values ranges}}$$

$$\underbrace{[0,1]\quad [0,1]\quad [0,1]\ \ [0,1]\ [0,1]\ [0,1]\ [0,1]\quad [0,1]}_{\text{attribute`s domains}}$$

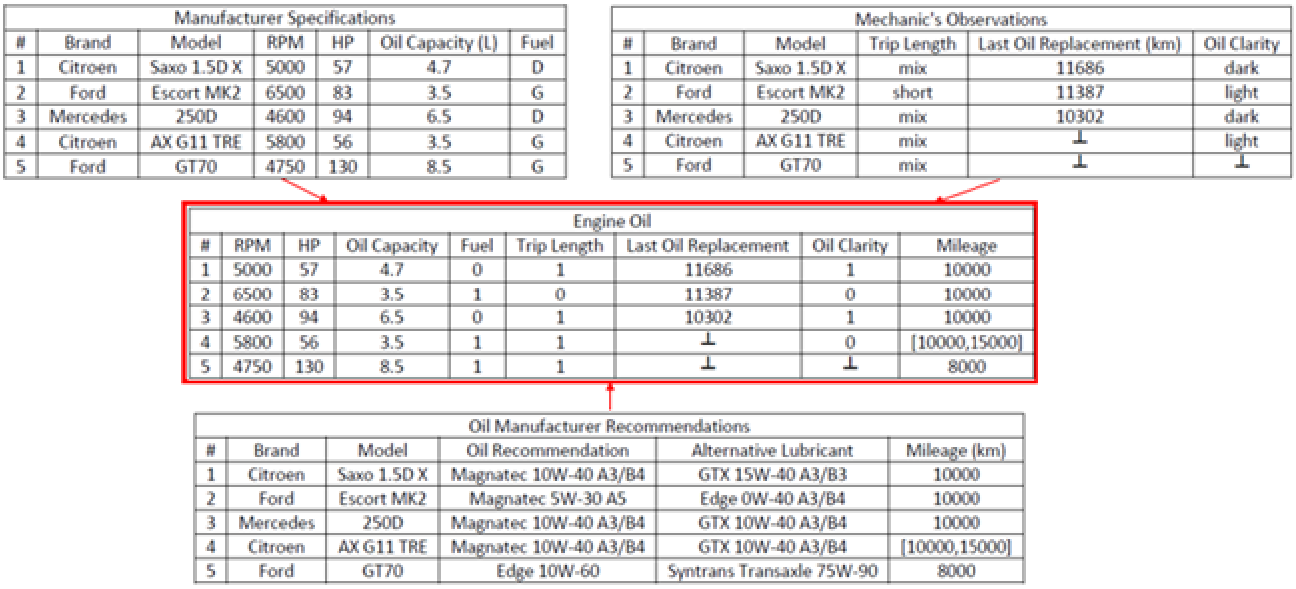
$$engine_{oil_{DoC}}(1,\quad 1,\quad 1,\quad 1,\quad 1,\quad 0,\quad 1,\quad 0.7)\quad ::\ 1$$

$$\underbrace{[0.632,0.632], [0,0], [0,0], [1,1], [1,1], [0,1], [0,0][0.286,1]}_{\text{attribute`s values ranges}}$$

$$\underbrace{[0,1]\quad [0,1]\ [0,1]\ [0,1]\ [0,1]\ [0,1]\ [0,1]\quad [0,1]}_{\text{attribute`s domains}}$$

$$\}$$

where its terms make the training and test sets of the Artificial Neural Network given below (Fig. 3).



Manufacturer Specifications						
#	Brand	Model	RPM	HP	Oil Capacity (L)	Fuel
1	Citroen	Saxo 1.5D X	5000	57	4.7	D
2	Ford	Escort MK2	6500	83	3.5	G
3	Mercedes	250D	4600	94	6.5	D
4	Citroen	AX G11 TRE	5800	56	3.5	G
5	Ford	GT70	4750	130	8.5	G

Mechanic's Observations					
#	Brand	Model	Trip Length	Last Oil Replacement (km)	Oil Clarity
1	Citroen	Saxo 1.5D X	mix	11686	dark
2	Ford	Escort MK2	short	11387	light
3	Mercedes	250D	mix	10302	dark
4	Citroen	AX G11 TRE	mix	⊥	light
5	Ford	GT70	mix	⊥	⊥

Engine Oil								
#	RPM	HP	Oil Capacity	Fuel	Trip Length	Last Oil Replacement	Oil Clarity	Mileage
1	5000	57	4.7	0	1	11686	1	10000
2	6500	83	3.5	1	0	11387	0	10000
3	4600	94	6.5	0	1	10302	1	10000
4	5800	56	3.5	1	1	⊥	0	[10000,15000]
5	4750	130	8.5	1	1	⊥	⊥	8000

Oil Manufacturer Recommendations					
#	Brand	Model	Oil Recommendation	Alternative Lubricant	Mileage (km)
1	Citroen	Saxo 1.5D X	Magnatec 10W-40 A3/B4	GTX 15W-40 A3/B3	10000
2	Ford	Escort MK2	Magnatec 5W-30 A5	Edge 0W-40 A3/B4	10000
3	Mercedes	250D	Magnatec 10W-40 A3/B4	GTX 10W-40 A3/B4	10000
4	Citroen	AX G11 TRE	Magnatec 10W-40 A3/B4	GTX 10W-40 A3/B4	[10000,15000]
5	Ford	GT70	Edge 10W-60	Syntrans Transaxle 75W-90	8000

Fig. 2 An extension of the relational database model. 0 (zero) and 1 (one), in column Fuel stand for, respectively, for diesel and petrol. In column Trip Length 0 (zero) and 1 (one) denote, respectively, short and mix. In column oil clarity 0 (zero) denotes *light* and 1 (one) denotes *dark*.

IV. Artificial Neural Networks

Neves et al [21]–[23] demonstrated how Artificial Neural Networks (ANNs) could be successfully used to model data and capture complex relationships between inputs and outputs. ANNs simulate the structure of the human brain being populated by multiple layers of neurons. As an example, let us consider the case where one may have a situation that may lead to an oil change, which is given in the form:

$$\{$$

$$engine_{oil} attributes: RPM, HP, Cap, F, Trip, Repl, Cla, M$$

$$\Downarrow$$

$$engine_{oil}([4800,5000],\ \perp,\ 4.5,\ 0,\ 0,\ 10564,\ 1,\ 10000)\quad ::\ 1$$
$$\underbrace{[4600,6500]\ [56,130][3.5,8.5]\ \ [0,1]\ \ [0,1][10302,11686][0,1]\ [8000,15000]}_{\text{attribute`s domains}}$$

$$\Downarrow\ \textit{1st interaction: transition to continuous intervals}$$

$$engine_{oil}([4800,5000], [56,130], [4.5,4.5], [0,0], [0,0], [10564,10564], [1,1], [10000,10000])\ ::\ 1$$
$$\underbrace{[4600,6500]\ \ [56,130]\ \ [3.5,8.5]\ [0,1]\ [0,1]\ \ [10302,11686]\ \ [0,1]\ \ \ [8000,15000]}_{\text{attribute`s domains}}$$

$$\Downarrow \textit{2nd interaction: normalization}\ \frac{Y - Y_{min}}{Y_{max} - Y_{min}}$$

$$engine_{oil}([0.105,0.211], [0,1], [0.2,0.2], [0,0], [0,0], [0.189,0.189], [1,1], [0.286,0.286])\ \ ::\ 1$$
$$\underbrace{[0,1]\qquad [0,1]\quad\ [0,1]\quad\ [0,1]\ \ [0,1]\qquad\ [0,1]\qquad\ [0,1]\qquad\ [0,1]}_{\text{attribute`s domains}}$$

$$\Downarrow\ \textit{DoC calculation: } DoC\ =\sqrt{1-\Delta l^2}$$

$$engine_{oil_{DoC}}(0.994,\quad 0,\quad 1,\quad 1,\quad 1,\quad 1,\quad 1,\quad 1) \quad :: \ 1$$

$$\underbrace{[0.105,0.211],[0,1],[0.2,0.2],[0,0],[0,0],[0.189,0.189],[1,1],[0.286,0.286]}_{\text{attribute`s values ranges}}$$

$$\underbrace{[0,1]\quad [0,1]\quad [0,1]\quad [0,1]\quad [0,1]\quad [0,1]\quad [0,1]\quad [0,1]}_{\text{attribute`s domains}}$$

}

In Fig. 3 it is shown how the normalized values of the interval boundaries and their *DoC* and *QoI* values work as inputs to the ANN. The output translates the chance of being necessary to go ahead with an oil change, and $engine_{DoC}$ the confidence that one has on such a happening. In addition, it also contributes to build a database of study cases that may be used to train and test the ANNs.

V. Future Work and Conclusions

To set a timeline to oil change is a hard and complex task, which needs to consider many different conditions with intricate relations among them. These characteristics put this problem into the area of problems that may be tackled by AI based methodologies and techniques to problem solving. Despite that, little to no work has been done in that direction. In this work it is 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 based in ANNs. This finding has several reasons, namely:

- Data do not equal to information;
- The translation of the raw measurements into interpretable and actionable read-outs is challenging; and
- Read-outs can deliver markers and targets candidates without pre-conception, i.e., knowing how motor conditions and risk factors may affect oil change.

Indeed, one´s approach to decision making is above everything else, very versatile and capable of covering every possible problem instance by considering incomplete, contradictory, and even unknown data. Future work may recommend that the same hitches has to be approached using others computational frameworks like Case Based Reasoning or Particle Swarm, just to name a few.

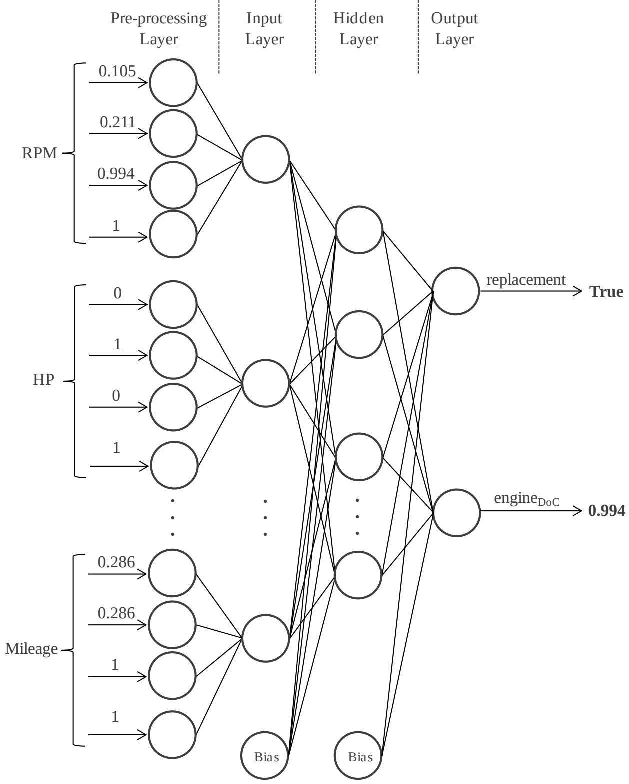


Fig. 3 The Artificial Neural Network topology

References

[1] S. Schwartz and D. Smolenski, “Development of an Automatic Engine Oil-Change Indicator System” SAE Technical Paper 870403, 1 February 1987.

[2] E. S. Schwartz, D. J. Smolenski, J. D. Keersmaekers, C. M. Traylor and G. J. Wallo, “Automatic engine oil change indicator system”, U.S. Patent 4 847 768, July 11, 1989.

[3] T. Sawatari, M. Nakamura, and T. Sugiura, “Automotive engine oil monitoring system”, U.S. Patent 4 677 847, July 7, 1987.

[4] B. Oliver, E. Cosgrove and J. Carey, “Effect of suspended sediments on the photolysis of organics in water”, *Environmental Science & Technology*, Vol. 13, pp. 1075-1077, 1979.

[5] J. Sarangapani, “Method for determining the condition of engine oil based on TBN modeling”, U.S. Patent 5 987 976, November 23, 1999.

[6] Castrol Limited. (2014, June 6) Castrol Portugal – Qual o lubrificante que devo utilizar? [Online]. Available: http://www.castrol.com/castrol/sectionbodycopy.do?categoryId=9006774&contentId=7077182.

[7] J. L. Vajgart, P. W. Misangyi, P. G. Date, R. S. Heitzeg, N. A. Walker, D. A. McNamara and J. C. Attard, “Method and apparatus for determining engine oil change intervals according to actual engine use”, U.S. Patent 5 060 156, October 22, 1991.

[8] E. T. King, “Method and apparatus for determining excessive engine oil usage”, U.S. Patent 4 970 492, November 13, 1990.

[9] J. C. Wang, S. D. Whitacre, M. L. Schneider and D. H. Dringenburg, “System and method for determining oil change interval”, U.S. Patent 6 253 601, July 3, 2000.

[10] H.-B. Jun, D. Kiritsis, M. Gambera and P. Xirouchakis, “Predictive algorithm to determine the suitable time to change automotive engine oil”, *Computers & Industrial Engineering*, Vol. 51, pp. 671-683, 2006.

[11] A. Kakas, R. Kowalski and F. Toni "The role of abduction in logic programming", in *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 5, D. Gabbay, C. Hogger and I. Robinson, Eds., Oxford: Oxford University Press, 1998, pp. 235-324.

[12] M. Gelfond and V. Lifschitz, "The stable model semantics for logic programming", in *Logic Programming – Proceedings of the Fifth International Conference and Symposium*, R. Kowalski and K. Bowen, Eds. Cambridge: MIT Press, 1988, pp. 1070-1080.

[13] L. Pereira and H. Anh, "Evolution prospection", in *New Advances in Intelligent Decision Technologies – Results of the First KES International Symposium IDT 2009*, K. Nakamatsu, G. Phillips-Wren, L. Jain and R. Howlett Eds. Studies in Computational Intelligence, vol. 199, Berlin: Springer, 2009, pp. 51-64.

[14] J. Neves, J. Machado, C. Analide, A. Abelha and L. Brito, "The halt condition in genetic programming", in *Progress in Artificial Intelligence - Lecture Notes in Computer Science*, vol 4874, J. Neves, M. F. Santos and J. Machado Eds. Heidelberg: Springer, 2007, pp. 160-169.

[15] J. Neves, "A logic interpreter to handle time and negation in logic data bases", in *Proceedings of the 1984 annual conference of the ACM on the fifth generation challenge*, R. L. Muller and J. J. Pottmyer Eds. New York: Association for Computing Machinery, 1984, pp. 50-54.

[16] J. Halpern, *Reasoning about uncertainty*. Massachusetts: MIT Press, 2005.

[17] B. Kovalerchuck and G. Resconi, "Agent-based uncertainty logic network", in *Proceedings of the IEEE International Conference on Fuzzy Systems – FUZZ-IEEE 2010*, Barcelona, Spain, 2010, pp. 596-603.

[18] P. Lucas, "Quality checking of medical guidelines through logical abduction", in *Proceedings of AI-2003 (Research and Developments in Intelligent Systems XX)*, F. Coenen, A. Preece and A. Mackintosh, Eds. London: Springer, 2003, pp. 309-321.

[19] J. Machado, A. Abelha, P. Novais, J. Neves and J. Neves, "Quality of service in healthcare units", *International Journal of Computer Aided Engineering and Technology*, vol. 2, pp. 436-449, 2010.

[20] Y. Liu and M. Sun, "Fuzzy optimization BP neural network model for pavement performance assessment", in *Proceedings of the 2007 IEEE International Conference on Grey Systems and Intelligent Services*, Nanjing, China, 2007, pp. 18-20.

[21] A. T. Caldeira, M. R. Martins, M. J. Cabrita, C. Ambrósio, J. M. Arteiro, J. Neves and H. Vicente, "*Aroma Compounds Prevision using Artificial Neural Networks Influence of Newly Indigenous Saccharomyces SPP in White Wine Produced with Vitis Vinifera Cv Siria*", in *FOODSIM 2010*, V. Cadavez and D. Thiel Eds. Ghent: Eurosis – ETI Publication, 2010, pp. 33–40.

[22] H. Vicente, S. Dias, A. Fernandes, A. Abelha, J. Machado, and J. Neves, "Prediction of the Quality of Public Water Supply using Artificial Neural Networks", *Journal of Water Supply: Research and Technology – AQUA*, vol. 61, pp. 446-459, 2012.

[23] H. Vicente, J. C. Roseiro, J. M. Arteiro, J. Neves and A. T. Caldeira, "Prediction of bioactive compounds activity against wood contaminant fungi using artificial neural networks", *Canadian Journal of Forest Research*, vol. 43, pp. 985-992, 2013.

Quality of Experience of Video in Transmedia Interactive Application of Digital Terrestrial Television

Cristiane Zakimi Correia Pinto, and Wagner Luiz Zucchi

***Abstract*—**The Brazilian digital terrestrial television system adopted in 2006 yields more than sounds and images of better quality when compared to legacy analog broadcast system, also makes it possible to receive TV signals on portable devices and mobile television receivers besides enabling interactivity. This paper presents an analysis of Quality of Experience (QoE) that was made measuring the user perception with an interactive application of digital terrestrial television. Such an application was developed using NCL (Nested Context Language), which is the standard declarative programing language of Brazilian digital television system called *Integrated Services Digital Broadcasting - Terrestrial Brazil* (*ISDB-TB*). In that application a secondary video is loaded through a broadband Internet access simultaneously with the main video being received through broadcasting. A test platform was created where IP packet loss was introduced in a controlled way affecting the secondary video, as it is expected to occur in a real network. Video quality was assessed for each loss level with objective metrics in order to compare QoE in each situation.

***Keywords*—** Digital television, ISDB-TB, Quality of Experience, video signal processing.

I. Introduction

At the same time that transmedia productions are becoming more common, legal changes are occurring in Brazil that bring an increase on daily interactivity use: the growth of the broadband access to Internet [1] and the adoption of a digital terrestrial television system that, when compared to legacy analog broadcast system, yields more than sounds and images of better quality, but also makes it possible to receive TV signals on portable devices and mobile television receivers besides enabling interactivity.

Digital terrestrial television is the digital television broadcast over the air free of charge to all population. It is different from digital cable television or digital satellite television that are pay TV services.

In Brazil, it can be stated that television is a major media platform. According to data from the Brazilian Institute of Geography and Statistics [2], the television set is present in 96.9% of Brazilian households, what is more than the number of households with a refrigerator (95.8%). Only the stove is found in more homes than television set (98.6%).

Considering the number of potential users of digital terrestrial TV, the use of interactive applications including this media platform is an important aid to the execution of a transmedia narrative.

One of the possibilities in the Brazilian Digital Terrestrial Television standard is to use the broadband Internet as the main interactive channel [3]. Considering this possibility, the study of Quality of Experience (QoE) perceived was made based on the assessment of the video quality obtained when using an interactive transmedia application of digital terrestrial television. Such an application was developed using NCL (Nested Context Language), which is the standard declarative programing language of Brazilian digital television system. In that application a secondary video is loaded through a broadband Internet access simultaneously with the main video being received through broadcasting.

The use of a broadband Internet as the main interactive channel is relevant because the time of loading the application and its corresponding data is critical to synchronize the application and the broadcast programming.

A test platform was created where IP packet loss was introduced in a controlled way affecting the secondary video, as it is expected to occur in a real network. The tests were made with the premise that the main video does not suffer interference, keeping its original quality.

The quality of the videos obtained in each case of packet loss was analyzed according to objective metrics and the obtained data allowed to infer what would be the QoE perceived in each case.

II. Transmedia

There has been a lot of talk about multimedia and crossmedia. More recently, with the popularization of the Internet and the appearance of new technologies that facilitated the interaction between people (like social networks), the concept of crossmedia evolved to transmedia, as defined for the first time by Henry Jenkins in [4].

Multimedia is often define as the existence of more than one means of communication of the same content, just a copy of the content in different media (e.g., a soap opera that has been

C. Z. C. Pinto is with the Polytechnic School of University of São Paulo, mailing address: R. Sócrates No. 697 Ap. 22A, São Paulo, SP 04671-071 Brazil (phone: +55-11-5686-3022; e-mail: cristiane.zcp@gmail.com).

W. L. Zucchi is with the Polytechnic School of University of São Paulo, Av. Prof. Luciano Gualberto Trav. 3 No. 158 Sala C2-23, São Paulo, SP 05508-010 Brazil (e-mail: wzucchi@lps.usp.br).

produced and shown on television and then released in DVD and blu-ray) [5].

Crossmedia is the possibility of a content to appear in different means of communication to share some idea, but there is no connection between the parts of the story that are developed in each platform (e.g., a movie and a comic book with the same characters) [5].

Transmedia is the term used when a story unfolds across multiple media platforms with each new text making a distinctive and valuable contribution to the whole [4],[5].

The Matrix phenomenon is an example of transmedia: three movies (the first, "The Matrix", produced in 1999), a collection of short animation films detailing the Matrix universe ("Animatrix", produced in 2003), a series of comic books, and two games ("Enter the Matrix" video game, and a multiplayer online game set in the universe of "The Matrix"). In The Matrix, several situations that occur on a media platform are part of the story on another media platform.

Always when there is a transmedia narrative, each platform gives its contribution to the narrative, encouraging the engagement of the audience. The audience makes the connection between the platforms, contributing to the development of the story.

In Fig. 1, a visual comparison between multimedia, crossmedia, and transmedia connections is shown (adapted from [5]).

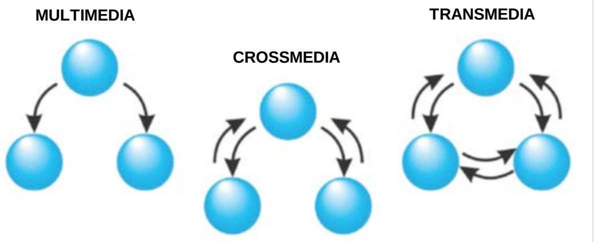


Fig. 1 multimedia, crossmedia, and transmedia connections

III. Digital Terrestrial Television

Digital terrestrial television is the digital television broadcast over the air free of charge to all population.

The differences between the analogue terrestrial television and digital terrestrial television are quite significant. In the case of Brazil, the digital television gives the possibility to receive high definition video and high quality audio, to receive TV signals on portable devices and mobile television receivers, to have more than one television program per channel and to have interactivity combined with the television programing.

A digital television programing is composed of a main audio and a main video, and may also contain additional data. This ability to transmit data, including applications relating the various media objects defined in these data, makes possible the offering of interactive services. The audio and video are delivered to the digital encoders that generate a main video stream and an audio stream, both compressed. These streams along with the data streams are multiplexed into one signal called the Transport Stream (TS). Then, the TS is modulated into a transmission channel, in frequency, and transmitted. At the reception, the signal is received and demodulated, being delivered to the demultiplexer which separates the main streams of audio and video from the data streams. The data streams are then delivered for processing. If used, the interactivity channel can provide new data to be processed [6].

A. Digital Television Systems

The digital television systems consist of a set of standards that make it possible processing, transmitting and receiving digital signals. Currently, in the world, there are the following digital TV systems: the American system ATSC (Advanced Television Systems Committee), the European system DVB-T (Digital Video Broadcasting - Terrestrial), the Japanese system ISDB-T (Integrated Services Digital Broadcasting - Terrestrial), the Chinese system DTMB (Digital Terrestrial Multimedia Broadcast) and the system adopted in Brazil ISDB-TB (Integrated Services Digital Broadcasting - Terrestrial Brazil), which is the Japanese system with modifications.

The choose of the ISDB-T system to be the base of the Brazilian system was made after a period of tests between the three existing systems at the time (ATSC, DVB-T and ISDB-T), in 2006 [7].

The Fig. 2 shows an overview of the architecture of the Brazilian digital terrestrial television system (ISDB-TB).

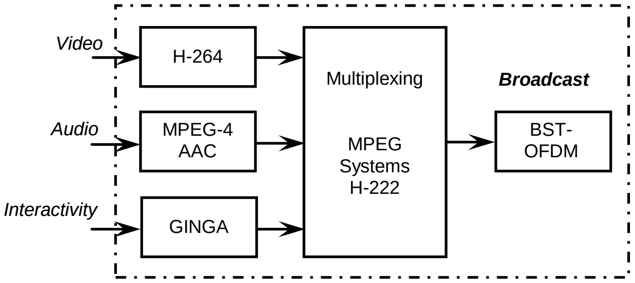


Fig. 2 overview of the ISDB-TB

B. Interactivity

The interactivity through digital television is different from the experienced when using the computer and accessing the Internet, for example.

A computer has large processing capacity and data storage and it is often operated by people who have a certain technical knowledge to install programs. On the other hand, the digital television receiver has limited resources, little memory, small or non-existent storage capacity, and the user cannot install components in the equipment. Therefore, the applications developed for interactive digital TV must consider the limitations of the system and be appropriate to the context.

It is important to clarify that the interactivity offered by digital terrestrial TV differs from the interactivity found in so-called "connected TVs". The "connected TVs" applications that provide access to interactive content (e.g., videos, tools for social networking, communication) are part of proprietary solutions that are the result of a partnership between content providers and manufacturers of television sets [8].

In the digital terrestrial television, the *middleware* is the responsible for the interactivity since it allows the applications and services can always be accessed in the same way, regardless of the platform they run on. In another words, the middleware is an intermediate layer between the common applications of digital TV system (such as electronic program guides and services offered by the TV station) and the applications that are defined by the receiver manufacturer, which is non-standard (such as the operating system used by the receiver).

Ginga is the middleware adopted in the ISDB-TB.

IV. Quality of Experience

The Quality of Experience (QoE) can be defined as a dependent parameter from the point of view of the user, which combines personal sensations and perceptions to judge whether the QoE is satisfactory or not [9], [10].

In this paper it was considered that the QoE is based on evaluation of video quality received when interacting with the application of digital TV.

V. Video Quality Assessment Methods

The video quality assessment can be made through subjective metric or through objective metric.

When a video is evaluated by subjective metrics, the judgment is based on human perception. The objective metrics use mathematical models to estimate user opinion.

The following objective metrics were used to evaluate the videos obtained by the digital interactive TV application for this paper [11], [12]:

A. PSNR (Peak Signal to Noise Ratio)

The PSNR defines the relationship between the maximum possible power of a signal and the noise that affects the representation of the signal between the frames of the original video and the degraded video:

$$PSNR = 10\log_{10} X$$
$$X = \frac{L^2}{\frac{1}{MNT}\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{t=1}^{T}\left[I(m,n,t)-\hat{I}(m,n,t)\right]^2}. \quad (1)$$

In (1), L is the is the dynamic range of pixel values, M and N are the width and the height of the original video and the degraded video sequences, respectively. T is the number of frames containing the sequences, *I(m,n,t)* and *Î(m,n,t)* represent the pixel in position *(m,n)* from *t* frame for the original sequence and the degraded video sequence, respectively.

B. SSIM (Structural SIMilarity)

This metric is based on the human visual model and it assumes that the images are highly structured, and these dependencies contain very important information relating to the structure of the object.

The SSIM algorithm estimates the similarity between the original video and the degraded video, comparing the brightness *l(x,y)*, contrast *c(x,y)* and structure *s(x,y)* of the original video *x* and the degraded video *y*.

The SSIM is represented by the following equation:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (2)$$

$$c_1 = (K_1 L)^2 \quad (3)$$

$$c_2 = (K_2 L)^2 \quad (4)$$

In (2), μ_x and μ_y is the average of x and y, σ_x^2 and σ_y^2 are the variance of x and y, respectively; σ_{xy} is the covariance of x and y; c_1 e c_2 are constants.

C. VQM (Video Quality Metric)

This method uses the Discrete Cosine Transform (DCT), and uses the original video and the degraded video as inputs to evaluate the video quality.

VI. Test Platform

The test platform used aimed to provide interactivity between television and computer platforms, which are interconnected with the TV station that broadcast the television programs. To do so, there are the following premises:

1) Television set has a digital TV signal receiver with interactivity, according to Brazilian standard [13].
2) Television set uses a broadband Internet as the main interactive channel.
3) The computer has access to broadband Internet to provide data to the interactive application.
4) There is no interference in the main video reception.
5) The user QoE is related only to the quality of the received video.

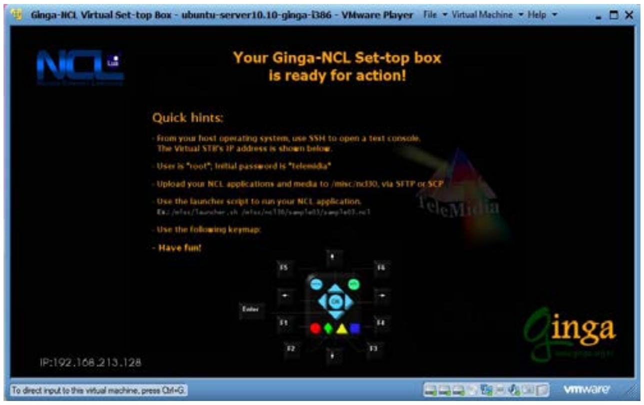
In the considered scenario, the TV station sends the interactive application along with the audio and video signals via data carousel [14], through broadcasting. The television set receives the data, displays on-screen programming and gives the possibility of interaction for the user (viewer). The user can interact and use the interactive channel (also called return channel), which is the broadband Internet.

To minimize the wait caused by loading data to compose the interactive application, the data will be accessed via Internet through the interactive channel rather than being transmitted over the air. Given the Brazilian standard, Internet content would be required by the application running [15].

With the described environment, the media platforms considered become interconnected, creating the necessary structure for the user to go through the various media and have a full transmedia experience.

To implement the test platform, the first step was the preparation of a transmedia interactive application that utilizes media platforms considered in this work. The application was written in NCL, which is mandatory for digital TV receivers.

The interactive application was used by GINGA-NCL digital TV set-top-box emulator, available free on the Internet [16].


Ginga-NCL Virtual Set-top Box - ubuntu-server10.10-ginga-i386 - VMware Player
File
Virtual Machine
Help
NCL
F5
F6
Enter
F1
F4
F2
F3
OK
ginga
vmware

Your Ginga-NCL Set-top box is ready for action!

Quick hints:

- From your host operating system, use SSH to open a text console. The Virtual STB's IP address is shown below.
- User is "root"; Initial password is "telemidia"
- Upload your NCL applications and media to /misc/ncl30, via SFTP or SCP
- Use the launcher script to run your NCL application. Ex: /misc/launcher.sh /misc/ncl30/sample03/sample03.ncl
- Use the following keymap:
- Have fun!

IP:192.168.213.128

To direct input to this virtual machine, press Ctrl+G.

Fig. 3 GINGA-NCL digital TV set-top-box emulator

Two computers were connected via UTP cable, where one of them was representing the *TV station* and the other was representing the *digital television set*.

The software WampServer [17] was installed in the TV station computer in order to transform it into a web server with the secondary video to be accessed. It also was installed the SoftPerfect Connection Emulator software [18] to introduce packet loss of the communication channel.

In the other computer that represented the digital television set the following software were installed: digital TV receiver emulator with interactivity GINGA-NCL [16], video recording software CamStudio [19], video editing software AVI Trimmer [20] and MSU Video Quality Measurement Tool [21].

The MSU Video Quality Measurement Tool determines the video quality with objective metrics. The tool is free to use in standard-definition videos. For high-resolution video, it is necessary purchase the PRO version license.

The main video was loaded into the computer that represented the digital television set because it was considered that there would be no interference in its reception.

With this test platform setting, it was possible to use the application of interactive digital TV, simulate various conditions of packet loss, and evaluate the quality of the received video.

VII. Results

The results obtained are shown in Table I.

Table I packet loss effect on the quality of the received video

Packet Loss (%)	PSNR (dB)	VQM	SSIM
1	34,20398	3,31409	0,98952
3	33,31370	4,20004	0,98885
5	33,23240	5,78433	0,98577
7	32,77781	6,22999	0,97684
10	30,77426	6,80946	0,97325

VIII. Conclusion

Analyzing obtained results, it appears that the quality of the resulting video for each situation of packet loss is as expected, i.e., the higher the packet loss, the worse the quality of the video displayed to the user and therefore worse QoE.

It also appears that the video quality in this study with the interactive application of digital TV resembles the case of receiving the video via streaming.

References

[1] Secretaria-Executiva do Comitê Gestor do Programa de Inclusão Digital. (2010). Programa Nacional de Banda Larga [Online]. Available: http://www.planalto.gov.br/brasilconectado

[2] G1 Economia. (2012, Sep. 20). Número de casas com TV supera o das que têm geladeira [Online]. Available: http://g1.globo.com/economia/noticia/2012/09/numero-de-casas-com-tv-supera-o-das-que-tem-geladeira.html

[3] *Digital Terrestrial Television - Interactive Channel. Part 1: Protocols, Physical Interfaces and Software Interfaces,* ABNT NBR 15607-1:2011 EN, 2011.

[4] H. Jenkins, *Cultura da Convergência*. São Paulo: Aleph, 2009.

[5] R. D. Arnault *et al.*, "Era Transmídia," *Revista GEMInIS*, ano 2 n.2, pp. 259–275, Jul./Dec. 2011.

[6] L. F. G. Soares, S. D. J. Barbosa, *Programando em NCL 3.0: Desenvolvimento de Aplicações para Middleware Ginga, TV Digital e Web*. Rio de Janeiro: Elsevier, 2009.

[7] Brazil, "Decreto n° 5.820 de 29 de junho de 2006," *Diário Oficial da União*, 2006, Jun. 30, pp. 51, 2006.

[8] A. C. B. Angeluci, R. D. Lopes, M. K. Zuffo. "Estudo comparativo entre TV digital aberta e TV conectada no Brasil," in *Proc. XXXIV Congresso Brasileiro de Ciências da Comunicação - INTERCOM*, Recife, 2011, pp. 1–16.

[9] R. Jain, "Quality of experience", *MultiMedia, IEEE*, vol. 11, no. 1, pp. 95–96, Jan.-Mar. 2004.

[10] Y. Lu, B. Fallica, F. A. Kuipers, R. E. Kooij, P. Van Mieghem, "Assessing the quality of experience of SopCast", *Int. J. Internet Protocol Technology*, vol. 4, no. 1, pp. 11–23, 2009.

[11] D. C. Begazo, D. Z. Rodríguez, M. A. Ramírez. (2011). Avaliação de qualidade de vídeo sobre uma rede IP usando métricas objetivas. *Revista Iberoamericana de Sistemas, Cibernética e Informática*, [Online]. Available: http://www.iiisci.org/journal/risci/Abstract.asp?var=&id=HCA940MZ

[12] D. Z. Rodríguez, G. Bressan., "Video quality assessments on digital TV and video streaming services using objective metrics". *Latin America Transactions, IEEE (Revista IEEE America Latina)*, vol. 10, no. 1, pp. 1184–1189, Jan. 2012.

[13] *Televisão Digital Terrestre – Receptores*, ABNT NBR 15604:2007 Versão corrigida:2008, 2008.

[14] *Digital Terrestrial Television - Data Coding and Transmission Specification Broadcasting. Part 3: Data transmission specification,* ABNT NBR 15606-3:2012 EN, 2012.

[15] *Digital Terrestrial Television - Data Coding and Transmission Specification Broadcasting. Part 1: Data coding specification*, ABNT NBR 15606-1:2013 EN, 2013.
[16] GINGA-NCL. [Online]. Available: http://www.gingancl.org.br/en
[17] Wampserver, a Windows web development environment. [Online]. Available: http://www.wampserver.com/en/
[18] SoftPerfect Connection Emulator. [Online]. Available: http://www.softperfect.com/products/connectionemulator/
[19] CamStudio. [Online]. Available: http://www.camstudio.org
[20] SOLVEIGMM. [Online]. Available: http://www.solveigmm.com/pt/products/avi-trimmer-mkv/
[21] VIDEO GROUP. [Online]. Available: http://compression.ru/video/quality_measure/video_measurement_tool_en.html

Cristiane Z. C. Pinto completed her Bachelor’s Degree in Electrical Engineering from the Mauá College of Engineering, Instituto Mauá de Tecnologia, São Caetano do Sul, SP, Brazil in 1996. Currently she is pursuing Master of Science degree in Electrical Engineering from Polytechnic School of University of São Paulo, São Paulo, SP, Brazil. Her areas of specialization include telecommunications and computer networks.

She is working as Telecommunications Engineer at the Diadema City Government, SP, Brazil since 2011. She has more than 19 years of working experience, mainly in the broadcast television industry. She and W. L. Zucchi have published the technical article “What is "Radio over fiber" and what are the applications of this technology?” in *RTI - Redes, Telecom e Instalações* (Brazilian monthly publication by Aranda Editors), vol. 120, pp. 92–95, 2010.

Engineer Pinto is a member of Brazilian Society of Television Engineering (SET - Sociedade Brasileira de Engenharia de Televisão) and Brazilian Society of Microwave and Optoelectronics (SBMO - Sociedade Brasileira de Micro-ondas e Optoeletrônica).

Wagner L. Zucchi received his Bachelor’s Degree in Electronic Engineering from the Polytechnic School of University of São Paulo, São Paulo, SP, Brazil in 1981 and Master of Science degree in Computer Networks from the same institution in 1989. He received his Ph.D. degree in the area of Computer Networks also from the Polytechnic School of University of São Paulo, São Paulo, SP, Brazil in the year 1998. His major field of study include Quality of Service in digital communication networks, network management, Data Center design and design of communication networks.

He is Professor at Electronic Systems Department of Polytechnic School of University of São Paulo, São Paulo, SP, Brazil since 1983 where he is responsible for undergraduate and graduate courses in the area of computer networks, performance evaluation and protocols. He has published technical papers in International Conferences and Journals, among them “Virtualization of Wireless Network Interfaces Wi-Fi IEEE 802.11.” with A. D. Rivera in *Proc. WSEAS International Conference on Telecommunications and Informatics*, Catania, 2010, pp. 46–51, and “Simulation of a optical burst switch using fiber delay lines” with M. C. F. Toledo, *IEEE Latin America Transactions*, vol. 6, pp. 28-34, 2008.

Prof. Dr. Zucchi is technical reviser of Information Systems and Computer Network books by Pearson editors in Brazil, such as *Computer Networking – A Top Down Approach* by Jim Kurose and Keith Ross, and *Structured Computer Organizations* by Andrew S. Tanenbaum. He is also monthly collaborator of *RTI – Redes, Telecom e Instalações* (*Networks, Telecommunications and Infrastructure*), Brazilian monthly publication by Aranda Editors.

Data warehouse minimization with ELT fuzzy filter

Jaroslav Zacek and Frantisek Hunka

Abstract**—** The paper proposes a new approach of data warehouse minimization by fuzzy-based ETL filter for ETL processes in business intelligence (BI) systems.

First part introduces common company systems and possible data sources in the company. Second part states the problem with interpreting information in BI systems and explains a data representation in the BI systems. Third part of the paper identifies suitable linguistic variables that help with interpreting the data to the user and automated filter as well. We also define a rule base and input and output values of expert system. Last part of the paper proposes a two ways to minimize a data - modification border of the fuzzy set and omitting useless combinations of the linguistic variables and modifiers.

Keywords**—** business intelligence, data minimization, linguistic variables, ETL process, expert system, ERP, data source.

I. Introduction

ALL companies are meant to produce a value. One of many tools describing the process of transforming the idea into value is called the business process. Business process is a collection of related activities to produce a specific product (or service). Therefore well-defined business processes are part of every successful company. Typical business process consists of events and activities. Activities can be supported by IT. For example the business process called “Create a car” can be composed probably of more then one activity and can be supported by many IT systems such as shipping system, supply chain system (SCM), economic agenda etc. All these systems are usually implemented as a separate unit and produce a specific data about their activities. Every business process should be measured to ensure optimal result for customer (in terms of quality, speed of delivery). One of the possibilities to measure the quality of the process is Key Performance Indicators (KPI). These indicators can be divided into several categories such as qualitative indicators, quantitative indicators, input or output indicators and can be very first management overview in the company.

KPI itself is not very sophisticated to making a management decisions in today’s information systems. According to Gartner predictions for the 2014 the digitalized supply chain approach will arise across companies deals with logistic problems. That means all supply strategies will not be only digitized, but also adaptive. Therefore the KPI cannot provide complex information. As we mentioned before there are usually more than one IT information system supporting the activities of the business process. These systems produce a large amount of multiple data and many companies’ uses a business intelligence (BI) to join, analyse and present desired data [7].

Business Intelligence (BI) is defined as a system to process large databases to support a decision-making in the company [5]. Sources of data can be internal (ERP system, SCM system) or external (mostly web services or automated web robots). Data can be processed from the distributed systems or from the internal database. BI has usually two components [1]. First component provide complex statistical software to analyse collected data. Second component is prediction core that estimates future trends. The purpose of BI is mainly to do a market statistical analysis and make decision based on customer and competition behaviour [5]. However, BI can be used to study and improve the business processes as well. BI is a tool naturally focused to larger companies, accessing large distributed databases. Some organizations have special department to tune up statistical calculations. However, small companies can also use a BI to understand, verify and improve the processes, which they manage. BI itself is not a new idea and has been considered since the beginning of business computing, during the time it has significantly developed. The practice has shown that a large database collecting a relevant data is more appropriate than service oriented data distributed data architecture. Therefore, we limit our hypothesis in this paper to database-oriented BI. This approach has another advantage – creating improved BI tools, which do not require IT experts to get benefits from large databases.

II. Problem formulation

There are three steps to realize business intelligence. First step is registration of the data source. This step finds a suitable data sources and prepares them to extract necessary information [6]. Data sources are very heterogeneous; we can utilize existing ERP or CRM system, web service, NoSQL databases or even a XML file. Every data source has specific format; for example data entity *User* can exist in both systems – in the ERP and same entity in the CRM system in conceptual way. However, an instance of the conceptual entity *User* has specific data types and code page according to database management system.

Second step is collecting data from external source and save data to one place. Data can be structured, unstructured or located into some analytic subsystem (i.e. OLAP). This part of

This work was supported by the European project OU OP VK (CZ.1.07/2.2.00/28.0245).

the architecture is crucial. A designer must select appropriate storage form because it affects speed of the whole system. Moreover, statistical methods are applicable in the second step. Computations of statistical data are time consuming therefore computation process runs in the background continuously. All data is pushed to data warehouse through ETL (extract, transform, load) processes. That processes can be imagine as a simple script (data pump), working in three steps:

- Extract – provides a simple data extraction from the data source.
- Transform – adjust data types, relations and prepares the data to the specific data architecture of the data warehouse.
- Load – pushes data into data warehouse.

ETL process can be time consuming; therefore these processes are running in parallelization [4].

When the data are collected, third step analyses collected data and presents data to the user (managers to support their decisions). Typical output of the BI process is a report in a form of table. If the user has another analytical systems BI can provide input data to these systems. There are also another possibilities how to present data from BI data warehouse. One of the favourite data representations is dashboard. It is a specific web page or part of regular application that provides an overview on all company data defined by user itself based on KPIs. Dashboards are limited to show summaries, trends and potential problems in business processes. To formulate the problem we have to introduce typical example in business process driven by ERP system.

Lets define a production company that uses ERP system to control the production and BI system to make decisions. All ERP systems strictly define parameters of every activity in the business process. Everything must be formalized because of invoice process and every purchase has a form of transaction. If the company orders 500 screws the transaction is not finished unless the specific amount is delivered. The transaction is not finished even if one screw is missing. From the point of ERP system view the situation has the justification. Order is connected with invoicing and warehouse management. But there is no reason to report that kind of information to the BI ecosystem.

First problem is how to detect and filter that kind of situation. In the real application could be significant count of "unfinished" orders in the system and BI evaluates this state as a serious problem and generate a false alarm in the dashboard. Second problem is where the filter should be applied. There are several places to implement filtration strategy – data source, ETL process, data warehouse, specific views and applications that show the results (dashboard, report).

Third problem is who defines and specifies filter parameters. For the purpose of explanation we limit the parameters to invoicing, but every hypothesis can be generalized.

Filter should be:

- Scalable – number of filter parameters depends on the variability of input data. Number of filter parameters can also be variable in time. Moreover common user must have a chance to change parameter by some specific GUI tool.
- Normalized – Filter will mix the specific parameters – for example quantity, date of delivery, price. All these parameters must be normalized before the decision-making function is applied.
- Adjustable by user – filter should be customized by common user in the simple way. Only user, not the implementer of the system, can declare importance of the specific attribute.

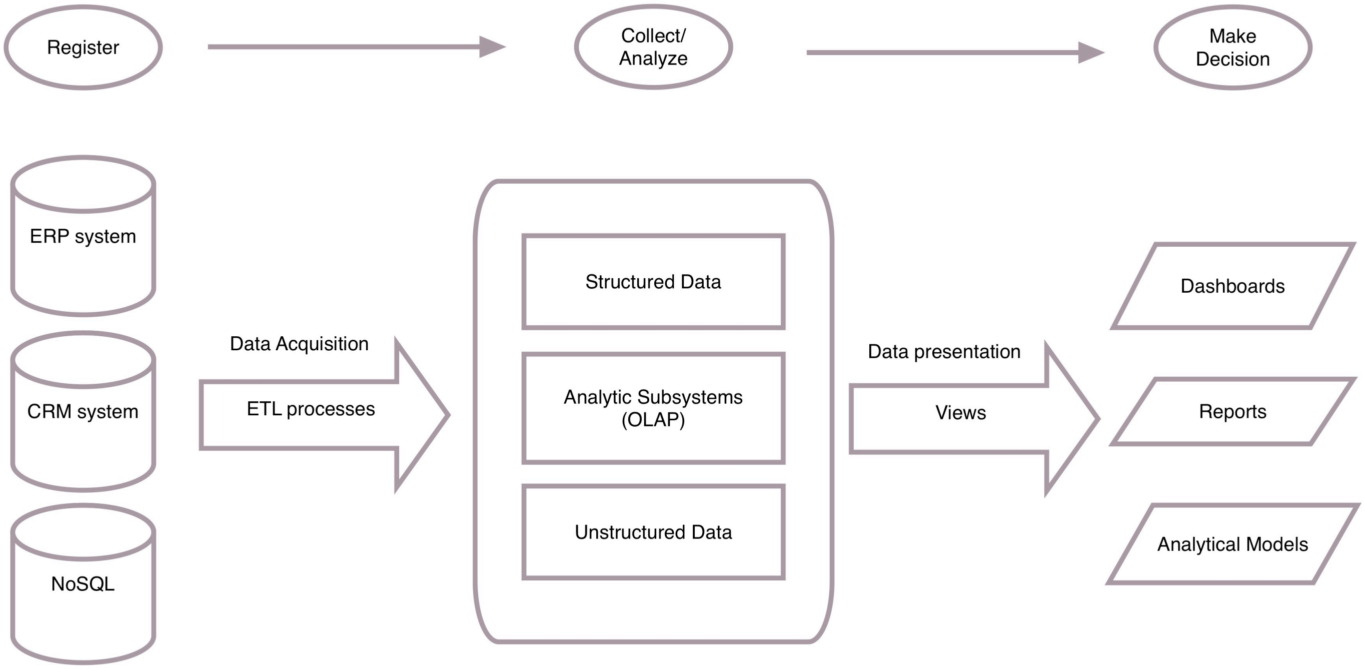


Fig. 1 General business intelligence architecture

All these requirements imply to use a fuzzy approach combining an expert system to support decision-making. Fuzzy approach enables to use linguistic variables to map discrete parameters to fuzzy sets and subsequently the rule base to make decision or ranking of the specific situation.

III. IDENTIFYING A SUITABLE LINGUISTIC VARIABLES

Linguistic variable is a term introduced by Lotfi A. Zadeh. He defined a linguistic variable as a linguistic expression instead of numeric value. Basic linguistic variables are defined in table 1 [2]. Linguistic variables are useful to describe parameter of the specific invoice. Common user cannot state that customer priority is 2, instead of the numeric value the user can state that customer priority is high.

Tab. 1: Basic linguistic variables

Linguistic variable	Abbreviation
Zero	ze
Small	sm
Medium	me
Big	bi

Basic linguistic variables cover the normalized interval <0;1>. However, this categorization does not have desired expressivity. Therefore we add modifiers to increase expressivity of the basic variables.

Tab. 2: Linguistic variables modifiers

Modifiers	Abbreviation
Very very roughly	vv
Typically	ty
Rather	ra
Very roughly	vr
Quite roughly	qr
Roughly	ro
More or less	ml
Very	ve
Significantly	si
Extreme	ex

By combination of basic linguistic variables and modifier we can assign specific linguistic variable to every parameter related to invoice. Every parameter domain of the filter can be described this way: shipping delay is very small; customer priority is roughly big; difference between ordered and delivered item is significantly small. Coverage of linguistic variables is show in Fig. 2. X-axis is a normalized domain interval <0,1> and Y-axis is a membership to the fuzzy set. Fig. 2 also highlights fuzzy set A that represents linguistic variable with modifier *more or less small* with kernel Ker(A) = 0.13, supremum Supp(A) = 0.23 and quadratic shape.

IV. DEFINE A RULE BASE OF THE EXPERT SYSTEM

Main part of the proposed filter is the expert system, which realizes a decision. The core of the expert system is the knowledge base containing the rules. Expert system uses fuzzy rules to make decisions regard to input linguistic values [3]. Fuzzy rules are the core concept of the fuzzy sets and fuzzy modelling application. Most commonly used fuzzy IF-THEN rule is written in specific form:

$$\mathcal{R} := \text{IF } \mathcal{X} \text{ is } \mathcal{A} \text{ THEN } \mathcal{Y} \text{ is } \mathcal{B}$$

Based on this form we can establish a rule base and perform a logic deduction:

$$\mathcal{R}_1 := \text{IF } (x \in A_1) \text{ THEN } (y \in B_1)$$
$$\text{AND}$$
$$\mathcal{R}_2 := \text{IF } (x \in A_2) \text{ THEN } (y \in B_2)$$
$$\text{AND}$$
$$\mathcal{R}_3 := \text{IF } (x \in A_3) \text{ THEN } (y \in B_3)$$

By Substituting x with linguistic variables we make a specific statement for one expert decision:

IF (customer priority is big) AND (shipping delay is very small) THEN (impact is very big)

IF (customer priority is small) AND (shipping delay is very big) THEN (impact is very small)

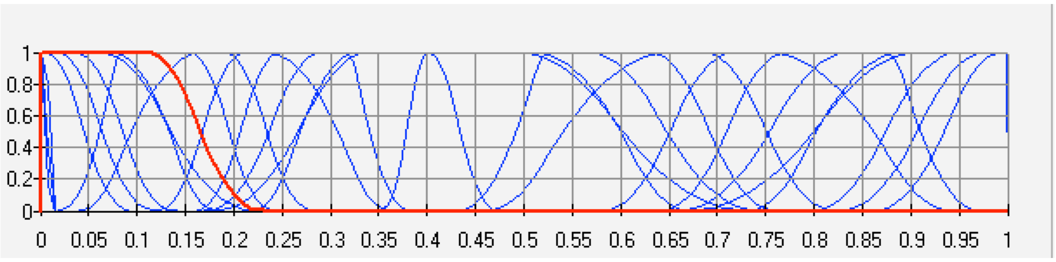
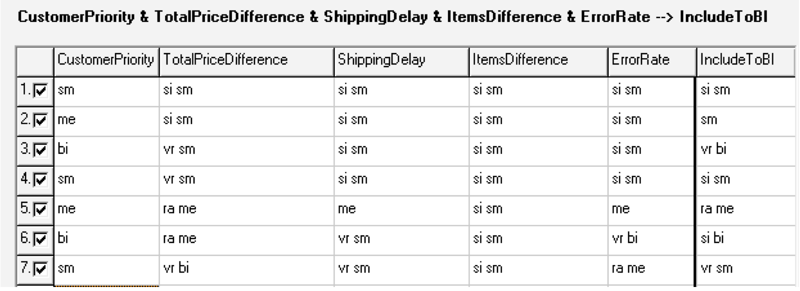


Fig. 2: Covering normalized intervals with fuzzy sets



CustomerPriority & TotalPriceDifference & ShippingDelay & ItemsDifference & ErrorRate --> IncludeToBI

	CustomerPriority	TotalPriceDifference	ShippingDelay	ItemsDifference	ErrorRate	IncludeToBI
1. ☑	sm	si sm	si sm	si sm	si sm	si sm
2. ☑	me	si sm	si sm	si sm	si sm	sm
3. ☑	bi	vr sm	si sm	si sm	si sm	vr bi
4. ☑	sm	vr sm	si sm	si sm	si sm	si sm
5. ☑	me	ra me	me	si sm	me	ra me
6. ☑	bi	ra me	vr sm	si sm	vr bi	si bi
7. ☑	sm	vr bi	vr sm	si sm	ra me	vr sm

Fig. 3: Rule base

The input values of our expert system are:

- *CustomerPriority* – determines a value of the customer for the company, linguistic values <sm; me; bi>
- *TotalPriceDifference* – represents difference between committed amount and paid amount, linguistic values <si sm; sm; vr sm; ra me; vr bi; bi; si bi>
- *ShippingDelay* – specifies difference between shipping date on the invoice and real delivery date, linguistic values <si sm; sm; vr sm; ra me; vr bi; bi; si bi>
- *ItemsDifference* – represents a difference between ordered item and delivered item, linguistic values <si sm; sm; vr sm; ra me; vr bi; bi; si bi>
- *ErrorRate* – indicates unclassified problems during ordering linguistic values <si sm; sm; vr sm; ra me; vr bi; bi; si bi>

The output of the expert system is only one – *IncludeToBI* variable with linguistic values <si sm; sm; vr sm; ra me; vr bi; bi; si bi>.

Rule base has been created in LFLC (Linguistic Fuzzy Logic Controller) developed at Institute for Research and Applications of Fuzzy Modelling at University of Ostrava. Decision-making mechanism affects type of inference and type of defuzzification. For this experiment we chose a perception-based logical deduction as an inference type and simple defuzzification of linguistic expressions (DEE). The rule base created in the LFLC tool is show in Fig. 3.

V. Connecting filter with BI

Business intelligence has three basic components and two transformation layers as seen in Fig. 1. Therefore we had five options where place proposed filtering mechanism and every choice has pros and cons. First option is to integrate filter directly inside data sources. Advantage of this choice is that proposed filter can be used inside the application for user reports (for example ERP delivery dates reports). This approach also saves significant computer time by filtering

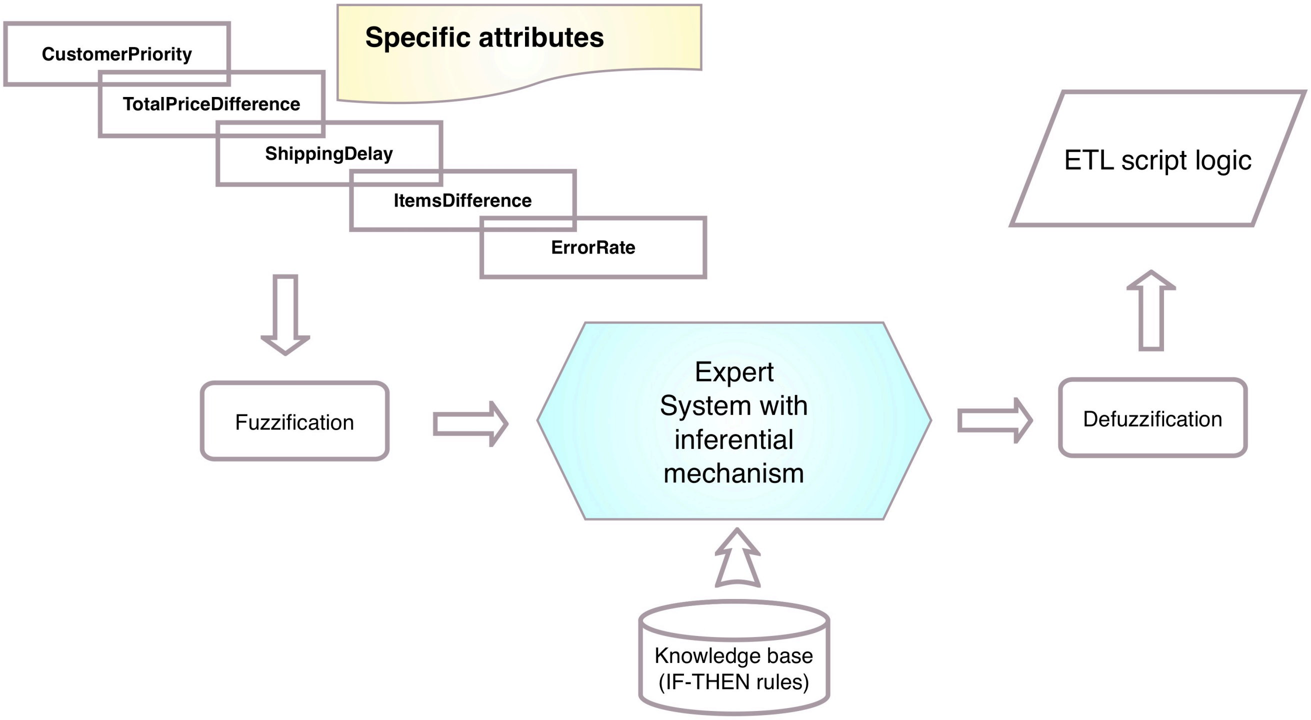


Fig. 4: Fuzzy-based filter architecture

result before the results are stored in the data warehouse. Disadvantage of this solution is that data sources are very heterogeneous and not every data source must be necessarily an information system. When the data source changes inner structure, fuzzy-based filter must be re-implemented.

Second option is to integrate filter directly into ETL Process. This approach saves significant computer time as well as the first option. Moreover it is independent on data source implementation. Every ETL process is connected to fuzzy-based filter; pushes input variables and make transform and load process based on expert system decision. First and second approach has one shared disadvantage – data are filtered before data warehouse with OLAP technology. That means if the user changes the rule base, change takes to effect after longer period (depends on the OLAP settings). Despite to the fact that we have significant delay by changing rule base, second option seems to be a good compromise.

Third option is to integrate filter directly into data warehouse implementation. This probably cannot be realized easily because of the proprietary implementation of the database management system (DBMS). There is no plugin option to enhance DBMS by external fuzzy-based decision system.

Fourth and fifth options are very common – implement fuzzy-based filter into graphical user interface (GUI), dashboard or other analytical model. Major advantage of this approach is that user gets immediate response by changing the rule base. Also in the case of fifth option user can manage linguistic variables directly in the GUI. The problem is that data warehouse analyses and stores irrelevant information and fuzzy-based filter must be implemented on every view.

Based on the facts described above we decided to put filter directly into ETL processes. Knowledge base is available to access by Web Services to edit IF-THEN rules from other part of the ecosystem (most likely dashboard or generic GUI). Proposed architecture including fuzzy-based filter is shown in Fig. 4.

VI. Data minimization

As we mentioned before, BI systems collect data from all available relevant data sources and put them into the data warehouse. This process can be limited by the size of the data storage system. Some data sources can be limited to specific number of queries per one day. First of all we have to find a way to minimize the amount of all data. BI systems analyze all collected data with statistical methods, which is extremely time and memory consuming.

First step to minimize data in the data warehouse is to exclude specific combination of linguistic variables and modifiers. Normally we stated about 10 modifiers and 4 linguistic variables and combination all of them should cover the normalized interval <0,1> with the fuzzy sets. Some combinations do not make sense in practice (for example Very Zero), some combinations extend the coverage of the interval and aggravate decision process (for example Significantly Medium). Therefore we will omit some combinations of linguistic variables and modifiers. Omitted combinations are summarized in table 3.

Tab. 3: Combinations that are not allowed

Modifier and hedge	Abbreviation
Typically Small	ty sm
Extremely Medium	ex me
Significantly Medium	si me
Very Medium	ve me
Typically Big	ty bi
Extremely Zero	ex ze
Significantly Zero	si ze
Very Zero	ve ze
Quite Roughly Zero	qr ze
Very Roughly Zero	vr ze
Rather Zero	ra ze
Typically Zero	ty ze
Very Very Roughly Zero	vv ze

Second step is to identify sets that represent naturally insignificant data and try to modify the border interval of the coverage. By this act can set a specific threshold to distinguish important data from unimportant without modification of the rule base.

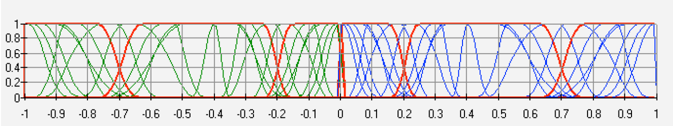


Fig. 5: Graphical representation of *roughly* modifier

VII. Conclusion

This paper proposed a fuzzy – based filter to minimize the size of the data warehouse used by BI systems. This approach can be used generally on any data that provides a data source into the data warehouse of the BI. Fuzzy approach makes filter scalable by user with linguistic variables such as small, medium and big.

Minimization process is realized by omitting useless combinations of modifiers and linguistic variables and by precision setting of the border interval of the fuzzy sets.

References

[1] Pour, J., Maryska, M., Novotny, O.: Business Intelligence in practice, Professional Publishing, (2012)

[2] Novak, V., Perfilieva, I.: Evaluating Linguistic Expressions and Functional Fuzzy Theories in Fuzzy Logic, In: Zadeh L.A., Kacpryk J. (eds.) Computing with Words in Information/Intelligent Systems 1, pp. 383-406. Springer-Verlag, Heidelberg (1999)

[3] Zadeh, L.A.: Fuzzy sets, Information & Control, vol. 8, pp. 338-353. (1965)

[4] Vandenbossche, P.E.A., Wortmann J.C.: Why accounting data models from research are not incorporated in ERP systems. In: Proceedings of the 2nd International REA Technology Workshop. (2006)

[5] McBride, N.: Business intelligence in magazine distribution, International Journal of Information Management, 34(1), 58-62 (2014)

[6] Hong, T.P., Yeong-Chyi L., Min-Thai W.: An effective parallel approach for genetic-fuzzy data mining, Expert Systems with Applications, 41(2), 655-662 (2014)

[7] Morgan, T.: Business Rules and Information Systems: Aligning IT with Business Goals, Addison-Wesley Professional (2002)

A platform for managing and forecasting certain aspects of the labor markets in Western Romania

K.B. Schebesch, C. Herman and A. Naaji

***Abstract*—**Understanding and managing modern labor markets poses some challenging problems for regional labor offices. The reasons and types of unemployment are manifold, and there exist various forms of very persistent as well as of hidden unemployment. In general, personal data about the unemployed is informative but it is also subject to some kind of privacy protection. A modern approach to managing local unemployment by labor offices is to be aware of all relevant current and past contextual evolutions of unemployment, to understand and forecast the entry of existing processes into and from different states of (un-) employment, and to optimize the usefulness of links between groups of unemployed people and effective labor market re-entry services. We created an online platform with an e-learning component which provides semi-automated procedures in order to solve such tasks.

***Keywords*—**online platform, e-learning, semi-automated procedures, labor market, unemployment, clustering and forecasting

I. Introduction

Modern labor markets pose some challenging problems for regional labor offices. Labor markets are associated with complex social processes and even the complete description of the relevant phenomena can pose serious problems. There are many kinds of employment types and various forms of unemployment. One may be concerned with explaining and predicting some aspects of labor markets or with developing measures to actively overcome some of the market dysfunctions. Labor relations and their dynamics may differ by degree of urbanity, by education and by region [1], [2]. Informal working relations are widespread over Europe and the world, for different aspects thereof see [3], [4], [5]. Entire categories of work may disappear [6], while new types of work and also of work-sharing may surface. However, the unrestricted use of the most informative personal data in order to find better ways of reacting to the various kinds of market mismatch may be subject to privacy constraints, an issue which often needs to be addressed by dedicated procedures [7]. Furthermore, labor markets are inducing nonlinear dynamic processes leading to somewhat counterintuitive social phenomena, such as “persistent inequality” [8] and other types of memory dependent nonlinear dynamical outcomes [9], [10]. The identification of regional employment [11] and of patterns in occupational mobility within and across regions or between job categories [12] is also of premier importance.

Finally, active labor market programs need to be developed [13]. The effects of such measures are often difficult to assess [14] which may be caused by the manifold overlapping and competing economic influences at any point in time. A possible way of relief from these is sketched by [1]. This by now famous article proposes multi-sided stable matching procedures, essentially a collective computational task which requires the design and development of a suitable problem-oriented platform, which we partially attempt to accomplish within an EU-financed labor market project (see acknowledgement).

II. Structure and Methodology of the Platform: an Application Set

The following application set is aimed at the need of unemployed people, namely to be informed and to return to work, and, for the employer, to recruit skilled workers. In a final stage the platform will provide at a basic level services such as: counseling and mediation, on-line interviews, help in curriculum vitae writing, personal and professional development plans, and how to apply for a job. The platform is a virtual meeting space between demand and supply and is operating in a (regional) labor market.

The platform uses SPOBDIA, a proprietary tool provided by partner ANALYTIKA, custom developed software for data entry and data analysis, as well as a combination modules developed under R 3.0.3 [15] and Scilab 5.5.0 [16] for the statistical modeling part of clustering and forecasting including some optimization and simulation tasks.

The software for data entry and data analysis was custom build to insure that we could reach the objectives of the project. Each module has an input/output functionality based

K.B. Schebesch is with the Department of Computer Science, “Vasile Goldi□” Western University, 310025 Arad, Romania (+40-257-285110; e-mail: kbeschebesch@uvvg.ro).

C. Herman is with the Department of Information Technology, “Vasile Goldi ”□ Western University, 310025 Arad, Romania (e-mail: cosmin@uvvg.ro).

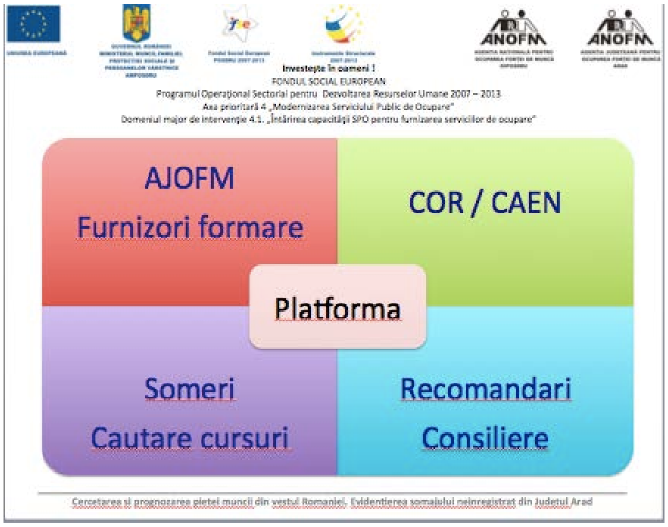
A. Naaji is with the Department of Computer Science, “Vasile Goldi ”□ Western University, 310025 Arad, Romania (+40-257-285813; e-mail: anaaji@uvvg.ro).

on data processing, gathering all the information in order to assure security and privacy of data.

The platform for online courses and recommendation system is based on the Moodle e-learning platform. All the integrations and application development was built around the need to have an online instrument for unemployed persons where they could search and find information about job opportunities, job enquiries and self assessment. The platform is an instrument to gather companies, unemployed people and Labor Agencies.

As depicted in Fig. 1 the platform is structured into four sectors or activity domains:

1. the County Labor Office;
2. the formal COR / CAEN structure (qualifications / related to economic activity domains) of courses offered online and offline;
3. a filter system for searching a specific course for a specific topic, and
4. a recommender system and personalized counseling. These features are built around an e-learning component which offers a well structured collection of courses.



Investește în oameni !
FONDUL SOCIAL EUROPEAN
Programul Operațional Sectorial pentru Dezvoltarea Resurselor Umane 2007 – 2013
Axa prioritară 4 „Modernizarea Serviciului Public de Ocupare"
Domeniul major de intervenție 4.1. „Întărirea capacității SPO pentru furnizarea serviciilor de ocupare"

AJOFM
Furnizori formare

COR / CAEN

Platforma

Someri
Cautare cursuri

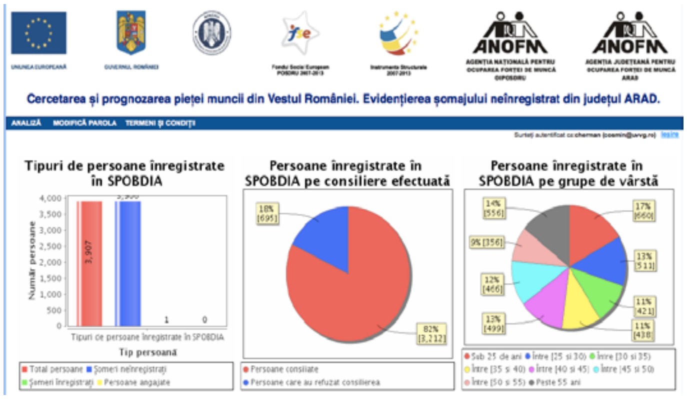
Recomandari
Consiliere

Cercetarea si prognozarea pietei muncii din vestul Romaniei. Evidentierea somajului neinregistrat din Judetul Arad

Fig. 1 the four activity domains or sub-processes of the integrated platform of the EU-funded project: suppliers of services (upper left), demand of unemployed persons (lower left), classification methodologies for qualifications, working experiences and jobs (upper right) and recommendations and counseling (lower right)

A. Data collection application interface

The existing databases, which hold personal (micro-) data about all temporarily registered unemployed persons in the three counties of the Western Romanian region (Arad, Bihor and Timis) are made available to the project partners. Additionally, the project partner ANALYTIKA (a firm specialized in labor market sociology) is engaged in collecting data by means of a rather complex and tedious process. Here special interview techniques are used to gather information about the "unregistered unemployed" (a category of hidden unemployment) in the region, and especially within the Arad county.



Cercetarea și prognozarea pieței muncii din Vestul României. Evidențierea șomajului neînregistrat din județul ARAD.

UNIUNEA EUROPEANĂ
GUVERNUL ROMÂNIEI
Fondul Social European POSDRU 2007-2013
Instrumente Structurale 2007-2013
ANOFM
AGENȚIA NAȚIONALĂ PENTRU OCUPAREA FORȚEI DE MUNCĂ OIPOSDRU
ANOFM
AGENȚIA JUDEȚEANĂ PENTRU OCUPAREA FORȚEI DE MUNCĂ ARAD
ANALIZĂ
MODIFICĂ PAROLA
TERMENI ȘI CONDIȚII
Sunteți autentificat ca cherman (cosmin@uvvg.ro) Ieșire
Tipuri de persoane înregistrate în SPOBDIA
Număr persoane
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
3,907
3,906
1
0
Tipuri de persoane înregistrate în SPOBDIA
Tip persoană
Total persoane
Șomeri neînregistrați
Șomeri înregistrați
Persoane angajate
Persoane înregistrate în SPOBDIA pe consiliere efectuată
18% [695]
82% [3,212]
Persoane consiliate
Persoane care au refuzat consilierea
Persoane înregistrate în SPOBDIA pe grupe de vârstă
14% [556]
17% [660]
9% [356]
13% [511]
12% [466]
11% [421]
13% [499]
11% [438]
Sub 25 de ani
Între [25 si 30)
Între [30 si 35)
Între [35 si 40)
Între [40 si 45)
Între [45 si 50)
Între [50 si 55)
Peste 55 ani

Fig. 2 data collection concerning the "unregistered unemployed"

This is part of the platform and is labeled SPOBDIA. The data collected here are much more detailed than those available from the existing data bases of the labor offices. In order to make them comparable some procedure of data fusion will be used.

The data collected is analyzed by a dedicated application in order to create raw data export. This data is then further used to generate groups (clusters) of unemployed persons and, finally, to forecasts the registered and the unregistered unemployed from the Arad county of Romania.

The components are:

- a *Data entry component,* i.e. reading a questionnaire applied to unemployed persons, thereby gathering various information about the status, location, education, work experience, health, future personal objectives, weather the person is looking for a job, whether the person is willing to change the location for a new job, etc.;

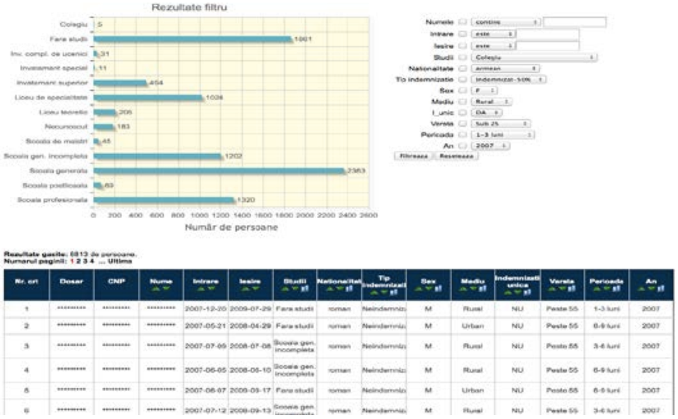
- a *Data-mining component* which analyzes and groups the information creating charts and clusters;

- an *Administration component* which is the zone where possible roles, professional abilities, effective work capacity and formal permissions of executing certain work by the candidates are being estimated or set; and finally,

- a *Data delivery component* (export), which is the part of the application that delivers structured data in order to be analyzed in more detail and by different means as will be described in section 3.

B. Managing some aspects of regional unemployment by using the data

As depicted in Fig. 3 one can use the databases from the local employment agencies which contain data of unemployed persons with entry and exit dates to the data base (4th and 5th column) and a list of personal characteristics (the columns to follow). A basic tool in managing un-employment on the basis of such information, which is typically spread out over several years, is to filter the records with regard to certain characteristics or levels of characteristics of unemployed (such a filter w.r.t. educational levels is shown at the upper part of Fig. 3). This leads to certain groupings of the unemployed based on "coordinate-wise" hypotheses (e.g. age-qualification groups, etc.) as would be advanced for instance by the experts of the labor offices. However, there are too many such potential groupings and it seems to be difficult to assess their usefulness of such groupings a priori.



Numele	□	contine
Intrare	□	este
Iesire	□	este
Studii	□	Colegiu
Nationalitate	□	armean
Tip indemnizatie	□	indemnizat-50%
Sex	□	F
Mediu	□	Rural
I_unic	□	DA
Varsta	□	Sub 25
Perioada	□	1-3 luni
An	□	2007

Rezultate gasite: 6813 de persoane.
Numarul paginii: 1 2 3 4 ... Ultima

Nr. crt	Dosar	CNP	Nume	Intrare	Iesire	Studii	Nationalitat	Tip indemnizati	Sex	Mediu	Indemnizati unica	Varsta	Perioada	An
1	*********	*********	*********	2007-12-20	2009-07-29	Fara studii	roman	Neindemniz	M	Rural	NU	Peste 55	1-3 luni	2007
2	*********	*********	*********	2007-05-21	2008-04-29	Fara studii	roman	Neindemniz	M	Urban	NU	Peste 55	6-9 luni	2007
3	*********	*********	*********	2007-07-09	2008-07-08	Scoala gen. incompleta	roman	Neindemniz	M	Rural	NU	Peste 55	3-6 luni	2007
4	*********	*********	*********	2007-06-05	2008-06-10	Scoala gen. incompleta	roman	Neindemniz	M	Rural	NU	Peste 55	6-9 luni	2007
5	*********	*********	*********	2007-06-07	2009-03-17	Fara studii	roman	Neindemniz	M	Urban	NU	Peste 55	6-9 luni	2007
6	*********	*********	*********	2007-07-12	2008-09-13	Scoala gen.	roman	Neindemniz	M	Rural	NU	Peste 55	3-6 luni	2007

Fig. 3 excerpt from a table of more than 50000 unemployed persons (micro-data) at the regional level of the Arad County

Note the confidential data entries of the persons from the first three columns. Using them freely or giving them away is certainly prohibited. However, they have a certain utility in determining, for instance, multiple entries in the database.

III. Forecasting unemployment and related clusters to labor market services

Using the micro data available from the databases of the labor office or, alternatively, using the micro data exported from the Data delivery component as described in section II B one may proceed to applying more refined clustering and forecasting techniques. Additionally one would be interested in defining some representation of active labor market programs along the lines of [13], [14].

Caracteristica: Nivel	Xtotal	Cluster 1	Cluster 2	Cluster 3	SPO 1
Var 3 STUDII : lev 1 Fara studii	8300	278	7863	159	0
Var 3 STUDII : lev 2 Scoala gen. incompleta	5225	563	3888	774	0
Var 3 STUDII : lev 3 Necunoscut	732	177	282	273	0
Var 3 STUDII : lev 4 Scoala generala	13711	5888	1548	6275	0
Var 3 STUDII : lev 5 Scoala profesionala	8031	2016	221	5794	0
Var 3 STUDII : lev 6 Liceu teoretic	1503	902	40	561	0.1
Var 3 STUDII : lev 7 Liceu de specialitate	8180	4342	52	3786	0.2
Var 3 STUDII : lev 8 Invatamant superior	4861	2857	30	1974	0.35
Var 3 STUDII : lev 9 Inv. compl. de ucenici	249	101	11	137	0
Var 3 STUDII : lev 10 Scoala de maistri	359	102	1	256	0
Var 3 STUDII : lev 11 Invatamant special	64	31	0	33	0
Var 3 STUDII : lev 12 Scoala postliceala	621	417	6	198	0.3
Var 3 STUDII : lev 13 Colegiu	40	20	1	19	0.05
Var 3 STUDII : lev 14 Invatamant superior f.l.	7	3	0	4	0
Var 3 STUDII : lev 15 Scoala speciala	8	6	0	2	0
Var 4 NAT : lev 1 roman	38959	16606	3412	18941	0
Var 4 NAT : lev 2 rrom	9921	119	9617	185	0
Var 4 NAT : lev 3 maghiar	2585	830	796	959	0
Var 4 NAT : lev 4 ceh sau slovac	194	87	49	58	0
Var 4 NAT : lev 5 german	111	30	26	55	0
Var 4 NAT : lev 6 necunoscut	37	8	13	16	0

Fig. 4 the 3rd and 4th variables from a typical table entry of the unemployed in the Western region of Romania (here Arad County, extracted over the years 2007-2012) are depicted by counting the number of unemployed persons with the corresponding levels of the respective variable (see main text)

Such programs are denoted by active labor market services (SPO) and can be thought of, for instances, as services to facilitate a group of unemployed persons to enter a new activity domain. Such a target group may contain, as a condition, at least 20% of unemployed persons of a certain age range, 30% from certain professions which are due to disappear from the local demand, etc. In such a way one can express each SPO by means of the same variables the single unemployed person. This is illustrated in Fig. 4. In addition we are grouping the unemployed into clusters which will hopefully separate the population into meaningful subgroups. For brevity we restrict the following description to a case of three clusters (remarks about the general case are found in the conclusions).

In Fig. 4 we confront the number of unemployed persons from the total population having a certain level of a characteristic (e.g. Var 3 / lev 5 "vocational school" for 8031 persons) in the total population with the corresponding number in the clusters C1-C3 (e.g. in C2 a total of 221 persons are found to have "vocational school" as their highest qualification) and with a desired shared of such person in a SPO-service (e.g. in our template SPO there is no requirement on a minimal number of people with "vocational school"). Continuing these tables in an automatic fashion for all variables/levels, for many more clusterings and for many more SPO-services (possibly to be implemented in parallel) we have a basis for a single objective or multi-objective optimization (assignment) problem (for the potential computational difficulty of these, consult e.g. [17] and [18]) of services to groups of sub-populations of unemployed persons. Next we would indicate how to further characterize the clusters obtained (by whatever statistical or human-proposed "qualitative" procedure). We propose to use to this end the description of the clusters by their implied time series in the spirit of [19]. Since clusters are sub-populations with individuals having entry and exit time flags attached, once we fix a suitable time resolution, we can compute times series as the number of entries minus the number of exits (in general not of the same persons) occurring during a time step. For this purpose we set a monthly time resolution (i.e. a time step that has the length of one month). Via these time series we facilitate direct comparisons of the clusters with the total population. Such time series are shown in Fig. 5.

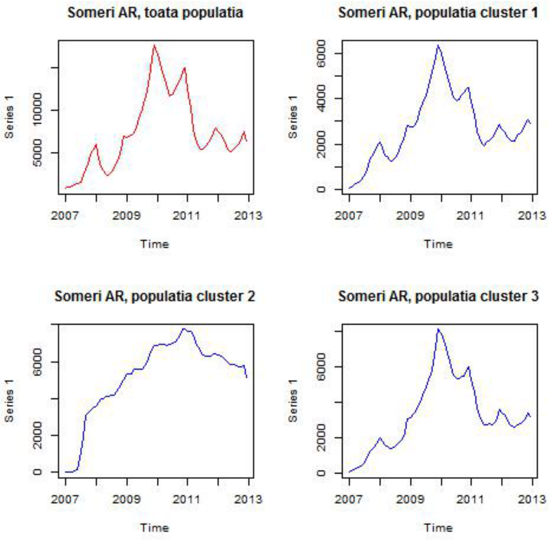


Fig. 5 time series (entries minus exists from the database) produced by the total unemployed population (upper left) versus time series produced by the populations of the respective clusters

One may observe that the time series generated by the populations in the clusters (the same three clusters indirectly described by the table entries from Fig. 4).

The information contained in Fig. 6 is an indicator of the forecastability of the respective time series. As the process is integrated time series [20], the "lag one" correlation is trivially positive and high (approaching 1). However, the remaining lags propose potentially useful linear forecasting models, if their correlations are located outside the confidence band (dashed lines at $\pm 1.96/\sqrt{N} \approx 0.23$, with N=72 entries or time steps).

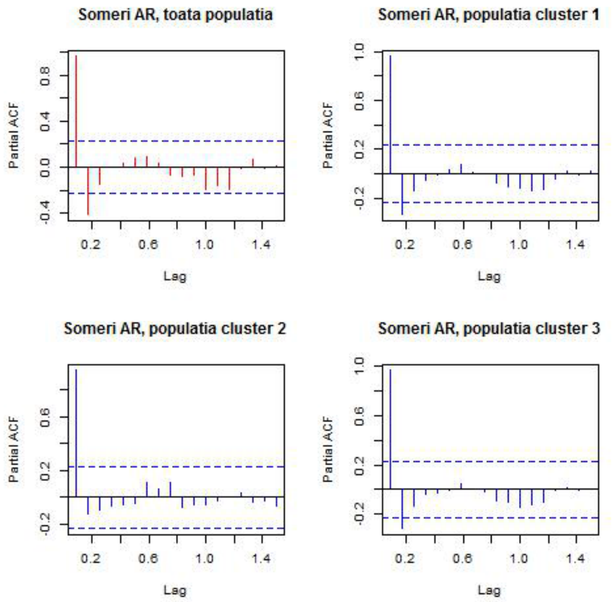


Fig. 6 the partial autocorrelation of the time series of the total population (upper left) and the time series extracted from the data of the clusters 1-3

The message of Fig. 6 is: compared to the time series of the original population, C1 and C3 may be forecasted with similar accuracy, while the more atypical cluster C2 is not forecastable – at least by linear models of the ARIMA class [20]. Furthermore, one may advance the hypothesis that the "unregistered unemployed" (which undeniably exist in substantial numbers in the area studied, but are not part of the data base) are similar to the population in C2 (to be confirmed by ongoing field work).

IV. Conclusions and Outlook

We explained and illustrated the usefulness of integrating several data-intensive processes in the context of labor market management and unemployment forecasting by regional labor offices. After describing and critically discussing various types of data collection procedures which occur in the context of regional (i.e. possibly in more specific types of unemployment), we proceed by using these data for different functional purposes as they would be required by a modern labor office. Starting out from the general concept for our platform we isolate and highlight two aspects, namely:

1. how to classify the ample micro-level data of registered unemployed persons into useful distinct groups of unemployed persons and how to characterize these groups or clusters by their implied time series of entries / exits over time, and

2. how to relate the clusters to the labor market services (SPO) to be offered directly or, indirectly, by using other pertinent service providers.

Although our examples presented here are limited to "small" (3-clusters and one SPO vector), in subsequent work (not shown here) we used clusterizations with as many as 30 simultaneous clusters of different emergent sizes. Upon analyzing these experiments we observe the following general pattern: Some time series can be forecast a) more easily, b) about as easy as and c) with much more difficult than forecasting the entire population. Clusters implying c) and (sometimes) a) may contain persons similar to the average hidden un-employed (the "unregistered" unemployed) of the region. However, in general none of the time series can be forecast well by using the popular ARIMA models, implying that some suitable nonlinear models [21], [22] may worthwhile alternative model candidates. Owing to the limited length of our data series (below 100 months), special care has to be taken in order to avoid overtraining. A further possibility which may in part alleviate this problem is using the different time series of the clusters as a kind of multivariate alternative in order to forecast the total population time series (or that of any particular cluster).

Next, our clusters can indeed be used in order to help assigning labor market services to (groups of) unemployed persons. Future interactive users of the platform (unemployed persons, mediators and employers, see [23]) should state their preferences with regard to as many aspects of supply and configuration of services, persons, groups of persons, etc., as possible. Fair solutions can then be found by using many-sided matching procedures as described in [24].

Acknowledgment

Part of this work was supported by the EU project direction 4.1, FSE project D.M.I. 4.1 – SOP HRD /111/4.1/S/91816. We acknowledge the cooperation of our project partners: the local/ regional labor offices (beneficiaries), the chambers of commerce, the labor market services and IT consultancy ANALYTIKA and the Regional Institute for Economic Planning of Tuscany (IRPET), Florence, Italy.

References

[1] R. Bénabou, "Workings of a city, location, education and production", in *Quarterly Journal of Economics*, 108:619-652, 1993.

[2] G.A. Christodoulakis and E.C. Mamatzakis, "Labour Market Dynamics in Greek Regions: a Bayesian Markov Chain Approach Using Proportions Data", in *Review of Economic Analysis* 2, pp.32-45 (*** different dynamics of regional employment ***), 2010.

[3] M. Hazans, "Informal Workers across Europe. Evidence from 30 European Countries", in *Policy Research Working Paper*, No 5912, World Bank, pp.22-39, December 2011.

[4] ILO, "*Labor Inspection in Europe: undeclared work, migration and trafficking*, Working Document No. 7, LAB/ADMIN, ILO, Geneva, 2010.

[5] F. Schneider and C. Williams, C., *The shadow economy*, Institute of Economic Affairs (IEA), Westminster, London, 2013.

[6] W.J. Wilson, "When Work Disappears: The World of the New Urban Poor", A.A. Knopf, Inc., 1996

[7] R. Stecking, R. and K.B. Schebesch, "Clustering for Data Privacy and Classification Tasks, in: Huisman, D. et al. (eds.), *Operations Research Proceedings 2013*, *Operations Research Proceedings*, DOI: 10.1007/978-3-319-07001-8_54, Springer International Publishing, Switzerland, 2014.

[8] B. Krauth, "A Dynamic Model of Job Networks and Persistent Inequality", in *Santa Fe Institute Working Papers*, SFI 1998-06-049, pp20, 1998.

[9] A.K. Misra and A.K. Singh, "A mathematical model for unemployment", in *Nonlinear Analysis: Real World Applications*, 12: 128–136, 2011.

[10] K.B. Schebesch and D. Deac, "Knowledge about replenishable resources: the dynamics of unemployment and job creation", in: Simian, D. (Ed.): Proceedings of the Third International Conference: *Modelling and Development of Intelligent Systems*, October 10-12, 2013, Sibiu, Romania, pp.119-126, 2014.

[11] P.C. Rotaru, "Empirical study on regional employment rate in Romania", in *Procedia - Social and Behavioral Sciences* 109, pp. 1365 – 1369, 2014.

[12] E.O. Lungu, A.M. Zamfir and C. Mocanu, "Patterns in the occupational mobility network of the higher education graduates. Comparative study in 12 EU countries", accessed by: arXiv:1306.5383v1 [physics.soc-ph] 23 Jun 2013

[13] J. Heckman, R.J. LaLonde and J.A. Smith, "The Economics and Econometrics of Active Labour Market Programs", in: Ashenfelter, O. and Card, D. (eds.), *The Handbook of Labour Economics*, Volume III.

[14] Carling, K. and Richardson K. (2004), The relative efficiency of labor market programs: Swedish experience from the 1990's, *Labour Economics*, 11(3), 335-54, 1999.

[15] R 3.0.3 for Windows at cran.r-project.org.

[16] Scilab 5.5.0 for Windows at www.scilab.org

[17] K. Maier, and V. Stix, "A Semi-Automated Approach for Structuring Multi Criteria Decision Problems", in *European Journal of Operational Research,* 225 (3) pp.487-496, 2013.

[18] A. Sinha, P. Malo, A. Frantsev and K. Deb, "Finding optimal strategies in multi-period multi-leader-follower stackelberg games using an evolutionary framework", *Computers and Operations Research*, London Amsterdam New York, Elsevier, 2013.

[19] B.D. Fulcher and N.S. Jones, "Highly comparative, feature-based time-series classification", accessed by: arXiv:1401.3531v1 [cs.LG] 15 Jan 2014

[20] J.D. Hamilton, "*Time Series Analysis*", Princeton, NJ: Princeton University Press, 1994

[21] M. Droumaguet, "Markov-Switching Vector Autoregressive Models: Monte Carlo experiment, impulse response analysis, and Granger-Causal analysis", PhD European University Institute, Florence, DOI 10.2870/63610, 2012.

[22] R.E. McCulloch and R.S. Tsay, "Statistical analysis of economic time series via Markov switching models", *Journal of Time Series Analysis,* 15, 523-539, 1994.

[23] Online labour exchanges: The workforce in the cloud, *The Economist*, June 1st 2013.

[24] Y. Shoham and K. Leyton-Brown, "*Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations"*, Cambridge University Press 2009 and 2010.

Visual Models Transformation in MetaLanguage System

Alexander O. Sukhov, Lyudmila N. Lyadova

***Abstract* —** At the process of creation and maintenance of information systems the model-based approach to the software development is increasingly used. This approach allows to move the focus from writing of the program code with using general purpose language to the models development with automatic generation of data structures and source code of applications. However at usage of this approach it is necessary to transform models constructed by various categories of users at different stages of system creation with usage of various modeling languages. An approach to models transformation in DSM platform MetaLanguage is considered. This approach allows fulfilling vertical and horizontal transformations of the designed models. The Metalanguage system support "model-text" and "model-model" types of transformations. The component of transformations is based on graph grammars described by production rules. Transformations of model in Entity-Relationship notation are presented as example.

***Keywords* —** Domain-specific languages, visual language, DSM platform, language workbench, model-based approach, model transformation, vertical transformation, horizontal transformation.

I. Introduction

Development of information systems with usage of the modern tools is based on the design of the various models describing the domain of the information system, defining data structures and algorithms of system functioning. The main idea of such model-driven approach is the systematic usage of models at various stages of software development that allows to shift the focus from writing code in general purpose programming language to building models and automatic generation of the source code and other necessary artifacts. At modeling developer abstracts from concrete technologies of implementations. It facilitates the creation, understanding and maintenance of models. This approach is intended to increase productivity and to reduce development time.

There are implementations of model-driven approach which use general purpose modeling languages for describing of information systems. So, the modeling language UML with the standard MOF (Meta-Object Facility) forms a basis of the concept MDA (Model-Driven Architecture) [1]. Other implementations of the model-driven approach are based on use of the visual *domain-specific modeling languages* (DSMLs, DSLs), intended to solve a particular class of problems in the specific domain. Unlike general purpose modeling languages, DSMLs are more expressive, simple in applying and easy to understand for different categories of users as they operate with domain terms. To support the process of development and maintenance of DSMLs the special type of software – *language workbench* (*DSM-platform*) – is used.

The various categories of specialists (programmers, system analysts, database designers, domain experts, business analysts, etc.) are involved in the process of information systems creation and maintenance. Often they need modification of modeling language description to customize and adapt DSML to new conditions, requests of business and possibilities of users. The transformations of models constructed by various users at different stages of information system creation with usage of various DSMLs are necessary for the models adjustment and integration [2].

For implementation of these possibilities it is necessary, that the language workbench allowed to build the whole hierarchy of models: model, metamodel, meta-metamodel, etc., where *model* is an abstract description on some formal language of system characteristics that are important from the point of view of the modeling purpose, a *metamodel* is a model of the language, which is used for models development, and a *meta-metamodel (metalanguage)* is a language for the metamodels description. Furthermore, the language workbench should contains the tools allowing to fulfill conversion of models between various levels of hierarchy (*vertical* transformations) and in one hierarchy level (*horizontal* transformations).

The *MetaLanguage* system is a language workbench for creating of visual dynamic adaptable domain-specific modeling languages. This system allows to fulfill multilevel and multi-language modeling of domain [3]. The basic elements of the metalanguage are *entity*, *relationship* and *constraint*.

Usage of domain-specific languages and tools for the system development also affects a transformation problem as there is a need of export of the models created with DSML to external systems which, as a rule, use one of the standard modeling languages that is different from used DSL. That is why one of the main components of the MetaLanguage system is the *transformer*. This component uses graph grammars for models transformations description. Implementation of graph grammars in the MetaLanguage system is defined by appointment of this language workbench.

This work was supported by Russian Foundation for Basic Research (grant 14-07-31330).

A. O. Sukhov is with the National Research University Higher School of Economics, Perm, Russia (phone: (+7) 912-589-0986; e-mail: Sukhov.psu@gmail.com).

L. N. Lyadova is with the National Research University Higher School of Economics, Perm, Russia (e-mail: LNLyadova@gmail.com).

II. Basic Concepts

The basic concept of transformation definition is a *production rule* which looks like $p : L \rightarrow R$, where *p* is a rule *name*, *L* is a *left-hand side* of the rule, also called the *pattern*, and *R* is a *right-hand side* of the rule, which is called the *replacement graph*. Rules are applied to the starting graph named the *host-graph*.

Let's suppose that four labeled graphs *G*, *H*, *L*, *R* are given, and graph *L* is a subgraph of graph *G*. Applying of the rule $p : L \rightarrow R$ to the starting graph *G* is called the replacement in graph *G* of subgraph *L* on graph *R*, which is a subgraph of graph *H*. The graph *H* is the result of this replacement [4].

Graph grammar is a pair *GG* = (*P*, G_0), where *P* is a set of production rules, G_0 is a starting graph of grammar.

Graph transformation is a sequenced applying to the starting labeled graph G_0 of finite set of rules $P = (p_1, p_2 \ldots p_n)$: $G_0 \overset{p_1}{\rightarrow} G_1 \overset{p_2}{\rightarrow} \ldots \overset{p_n}{\rightarrow} G_n$.

Transformations can be classified as horizontal and vertical according to direction. The *horizontal transformation* is the conversion, in which the source and target models belong to one hierarchy level. An example of a horizontal transformation is a conversion of model description from one notation to another (see Fig. 1). The *vertical transformation* converts the models which belong to various hierarchy levels, for example, at mapping of the metamodel objects to domain model objects.

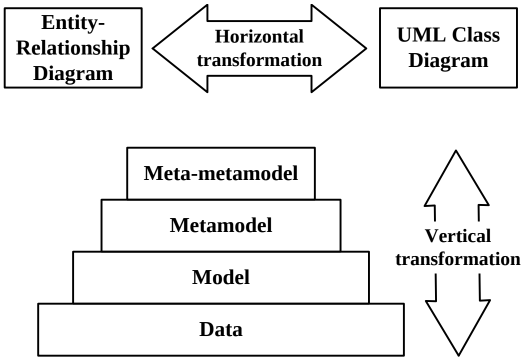


Fig. 1. Horizontal and vertical model transformations

The models are described with some modeling languages. Depending on the language on which source and target models are described, horizontal transformations can be divided into two types: endogenous and exogenous. An *endogenous transformation* is the transformation of the models, which are described on the same modeling language. An *exogenous transformation* is the transformation of models, which are described on various modeling languages [5].

Graph grammars are often used to describe any transformations performed on graphs: definition of the models operational semantics [6], the analysis of program systems with dynamic evolving structures [7], etc.

The right-hand side of the rule may be not only a labeled graph, but the code on any programming language, and also a fragment of a visual model described in some notation. That is why the graph grammar can be used for generation syntactic correct models and for refactoring of existing models, code generation and model transformation from one modeling language to another [8].

Considering singularities and designation of MetaLanguage system, it is necessary to make the following requirements to its transformation component:

- To be obvious and easy to use for providing the opportunity of involving to transformation description not only programmers, but also experts, specialists in domains. It can be achieved through the usage of visual notation of transformations description language.
- To allow using the created transformations directly in the system, i.e. to produce the models transformations in the same user interface, in which they were designed.
- To perform both horizontal and vertical transformations, and possibility to fulfill the horizontal transformations from one notation to another, including a “model-text” type.
- To allow specifying the transformations of entities and relationships attributes and constraints imposed on metamodel elements.

III. Related Works

There are various approaches to model transformations. Some of them have the formal basis, so the systems AGG, GReAT, VIATRA use graph rewriting rules to perform transformations, and others apply technologies from other areas of software engineering, for example the technique of programming by example.

Various modifications of the algebraic approach [9] are implemented in systems AGG, GReAT, VIATRA. In AGG (Attributed Graph Grammar) [10], [11] the left- and right-hand sides of the production rule are the typed attribute graphs, both sides of a rule should be described in one notation, i.e. this system allows to fulfill only endogenous transformations that does impossible its usage in MetaLanguage system. Besides, this tool does not allow to make transformation of a “model-text” type. However the usage as the formal basis of the algebraic approach to graph transformations allows to produce graph parsing, to verify graph models, and the extension of graphs of Java possibilities makes transformations more powerful.

The GReAT (Graph REwriting And Transformation) system [12], [13] is based on the algebraic approach with double-pushout, therefore for transformation description it is necessary to create the domain that contains both the left- and right-hand sides of the production rule simultaneously with instructions of what element it is necessary to add, and what to remove. This form of rule is unusual for the user and a bit tangled. However it provides a possibility of execution the transformation of several source metamodels at once, which is significant advantage in comparison with other approaches. For metamodels definition the GReAT uses UML and OCL, it does not allow the user to choose the language of metamodels specification or to change its description. It makes this approach unsuitable for usage in MetaLanguage.

The QVT (Query/View/Transformation) is the proposed by OMG approach to models transformation, which provides the user with declarative and imperative languages [14], [15]. Conversion is defined at the level of metamodels, which is described on MOF. The advantage of this approach is the existence of standard of its description, and also usage of standard languages OCL and MOF at the models transformation definition. But usage of MOF as a meta-metamodeling language, does not allow the user to choose a metalanguage convenient for him, or to change description of the metalanguage which is integrated in the QVT.

VIATRA (VIsual Automated model TRAnsformations) [16], [17] is a transformation language, based on rules and patterns, which combines two approaches into a single specification paradigm: the algebraic approach for models description and the abstract state machines intended for exposition of control flow. Thanks to constructions of state machines the developers significantly raised the semantics of standard languages of patterns definition and graph transformation. Besides, powerful metalanguage constructions allow to make multilevel modeling of domains. One of shortcomings of the VIATRA is an inexpressive textual language of metamodels description. VIATRA is not intended for execution of horizontal model transformations. Its main purpose is a verification and validation of the constructed models by their transformation.

The ATL (ATLAS Transformation Language) is the language, allowing to describe transformations of any source model to a target model [18], [19]. Transformation is performed at the level of the metamodels. The disadvantage of this language is high requirements to the developer of transformation. Since ATL in most cases uses only textual definition of transformation, then in addition to knowledge of source and target metamodels the developer needs to know language of transformation definition. The ATL is a dialect of QVT language and therefore inherits all its shortcomings.

MTBE (Model Transformation By-Example) approach [20], [21] is quite non-standard and unusual. The main purpose of MTBE is automatic generation of transformation rules on a basis of an initial set of learning examples. However implementations of this approach do not guarantee that the generation of model transformation rules is correct and complete. Moreover, the generated transformation rules strongly depend on an initial set of learning examples. Current implementations of MTBE approach allow to fulfill only full equivalent mappings of attributes, disregarding the complex conversions.

In summary, it is possible to say that all considered systems have some disadvantages which restrict their applicability for transformation definitions in the MetaLanguage system. But the most appropriate and perspective, from the author's point of view, is the algebraic approach.

IV. Model Transformations

Horizontal transformation is the conversion, in which the source and target models belong to one hierarchy level.

All horizontal transformations in MetaLanguage system are described at level of metamodels that allows to specify conversions which can be applied to all models created on basis of this metamodels. For a transformation definition it is necessary to select a source and target metamodels and to define production rules that are describing conversion.

To define the rule it is necessary to select objects (entities and relationships) in a source metamodel, to set constraints on pattern occurrence and to define the right-hand side of the rule. Depending on a type of transformation a right-hand side will be a text template for code generation, or a fragment of a target visual metamodel.

Transformation rules are applied according to their order. At first all occurrences of a first rule pattern will be found, for each of them the system will replace it by the right-hand side of the production rule, then the system will pass to the second rule and will begin to execute it, etc.

Let's assume that the system has selected next production rule of transformation and trying to execute it. For implementation of rule application it is necessary to describe two algorithms: the algorithm of the pattern search in the source host-graph and the algorithm of replacement of the left-hand side of the rule by the right-hand side.

There are various algorithms of search of subgraph isomorphic to the given pattern: Ullmann algorithm [22], Schmidt and Druffel algorithm [23], Vento and Foggia algorithm [24], Nauty-algorithm [25], etc. These algorithms are the most elaborated and often used in practice.

However difference of the proposed approach from the classical task of graph matching is that in this case it is necessary to find a pattern in the metamodel graph, i.e. it is required to lead matching of graphs which belong to various hierarchy levels, thus it is necessary to consider type of nodes and arcs, as between two nodes of the metamodel graph the several arcs of various type can be led.

The offered algorithm for finding a pattern in the graph model is a kind of backtracking algorithm that takes exponential time.

Since the amount of arcs in the model graph is less than amount of the nodes usually, each arc uniquely identifies nodes, that are incident to it, and the degree of node can be more than two, that does not allow to select the following node of the model graph, entering into a pattern. It was decided to start search of subgraph in a model graph on the basis of search of particular type arcs.

At the first step of algorithm all instances of some arbitrary relationship of the pattern will be found, i.e. search of an initial arc with which execution of the second step of algorithm will begin is carried out. At the second stage it is necessary to find one of possible occurrence of all relationships instances of the pattern-graph G_P in the source model graph G_S. At the third step necessary nodes will be add to target graph G_T and replace the left-hand side of the rule by the right-hand side.

Then it is necessary to replace the left-hand side of the production rule by the right-hand side after the subgraph of left-hand side has been found in the source graph. The algorithm of replacement will depend on a type of transformation: whether transformation is "model-text" or "model-model".

Transformation "model-text". The transformation of this type allows to generate the source code on any target programming language on the basis of the constructed models as well as any other textual representation of model, for example, its description on XML. In this case the right-hand side of production rule contains some template consisting of as static elements, which are independent of the found pattern, and dynamic parts, i.e. elements which vary depending of the found fragment of model.

For transformation fulfillment it is necessary to find all occurrences of a pattern in a source graph and to produce an insertion of an appropriate text fragment with a replacement of a dynamic part by appropriate names of entities, relationships, values of their attributes, etc.

The template is described on the target language. For selection of a dynamic part of a template the special metasymbols are used: "<<" (double opening angle brackets) to indicate the beginning of a dynamic part, ">>" (double closing angle brackets) to indicate the end of a dynamic part. As entities and relationships can have the same name, then for entity describing before its name the prefix "E." is specified, and for relationship describing before its name the prefix "R." is specified.

At the transformation specifying it is possible to set constraints on pattern occurrence. These constraints allow to define the context of the rule. They contain conditions with which found fragment of model should satisfy.

Let's consider an example: define the transformation that allows on the basis of Entity-Relationship Diagrams (ERD) to generate a SQL-query, building the schema of a corresponding database.

At the first step it is necessary to choose the metamodel of Entity-Relationship Diagrams (see Fig. 2) and to set the transformation rules.

The metamodel contains the entities "Abstract", "Attribute", "Entity", "Relationship". Attributes of the entity "Abstract" are "Name" that identifies an entity instance, and "Description", containing the additional information about the entity. The entity "Abstract" is abstract, i.e. it is impossible to create instances of this entity in the model. "Abstract" acts as a parent for entities "Entity" and "Relationship" (in the figure it is shown by an arrow with a triangular end). Both child entities inherit all parent attributes, relationships, constraints. "Entity" does not have own attributes and constraints. "Relationship" has the own attribute "Multiplicity". The entity "Attribute" has following attributes: "Name", "Type" and "Description".

The bidirectional association "Linked_Links" connects entities "Relationship" and "Entity". It means that it is possible to draw equivalent relationship between these entity instances in ERD-models. The second unidirectional association "SuperClass_SubClass" binds entity "Entity" with itself, it allows any instance of "Entity" to have parent (another instance of "Entity") in ERD-models. In ERD metamodel between entities "Attribute" and "Abstract" the aggregation "Belongs" is set (in figure this relationship is presented by an arc with a diamond end), therefore in ERD-models instances of entities "Relationship" and "Entity" can be connected by aggregation with the instances of entity "Attribute".

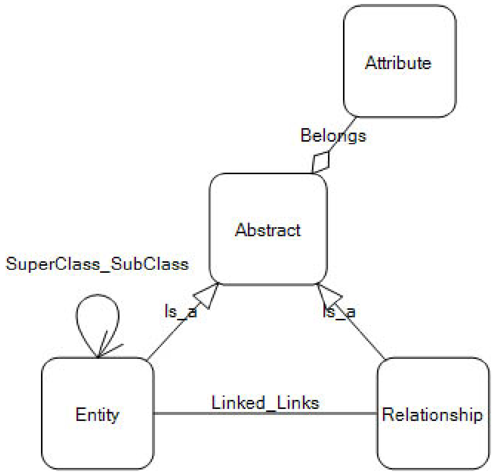
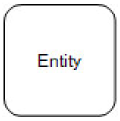


Fig. 2. Metamodel of Entity-Relationship Diagrams

For correct transformation execution the additional attributes in the source metamodel should be added. To determine what entity is a parent, and what entity is a child it is necessary to add the mandatory attributes of a reference type ("Child" and "Parent") to relationship "SuperClass_SubClass". The entity "Relationship" should be transformed to the reference between relational tables, therefore we will add to "Relationship" additional mandatory attributes-references of "LeftEntity" and "RightEntity" and attribute of logical type "Has_Attribute", which will facilitate the replacement of the left-hand side of the production rule by the right-hand side.

For transformation definition we will use the traditional rules of conversion of the ERD notation to a relational model, for this purpose we will define the following rules.

The rule "Entity" which transforms the instance of entity "Entity" to the single table looks like:



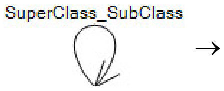
→

```
CREATE TABLE <<E.Entity.Name>>
(id INTEGER primary key)
```

Here `<<E.Entity.Name>>` is a dynamic part of the template which allows to get a name of corresponding entity.

As there is not inheritance relationship in a relational model, it is necessary to specify the rule "Inheritance", which for each instance of the relationship "SuperClass_SubClass" in the "SubClass" table creates foreign key for connection with the "SuperClass" table.

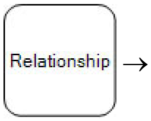
This rule looks like:



```
ALTER TABLE
<<R.SuperClass_SubClass.Child>>
ADD <<R.SuperClass_SubClass.Parent>>
ID INTEGER
ALTER TABLE
<<R.SuperClass_SubClass.Child>>
ADD FOREIGN KEY
(<<R.SuperClass_SubClass.Parent>>ID)
REFERENCES
<<R.SuperClass_SubClass.Parent>> (id)
```

The rule “Relationship_1M” allows to transform instance of entity “Relationship”, which does not have attributes and its multiplicity is “1:M”, to the reference between tables.

The rule has the following appearance:



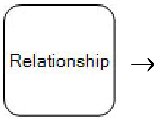
```
ALTER TABLE <<E.Relationship.LeftEntity>>
ADD <<E.Relationship.RightEntity>>
ID INTEGER
ALTER TABLE <<E.Relationship.LeftEntity>>
ADD FOREIGN KEY
(<<E.Relationship.RightEntity>>ID)
REFERENCES <<E.Relationship.RightEntity>>
(id)
```

In this rule at first in the table corresponding to the left entity the additional column with the name `<<E.Relationship.RightEntity>>ID` is added, and then the foreign key (correspondence between this additional column and a column containing the identifiers of right table rows) is created. This rule contains the constraint on the pattern occurrence:

```
E.Relationship.Multiplicity = 1:M AND
 E.Relationship.Has_Attribute = False
```

The rule “Relationship_M1” allows to transform instance of entity “Relationship”, which does not have attributes and its multiplicity is “M:1”, to the reference between tables.

The rule looks like:

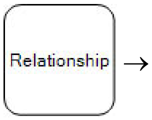


```
ALTER TABLE <<E.Relationship.RightEntity>>
ADD <<E.Relationship.LeftEntity>>
ID INTEGER
ALTER TABLE
<<E.Entity.Relationship.RightEntity>>
ADD FOREIGN KEY
(<<E.Relationship.LeftEntity>>ID)
REFERENCES
<<E.Relationship.LeftEntity>>(id)
```

The content of this rule is similar to the content of the rule “Relationship_1M”. This rule contains the constraint on the pattern occurrence:

```
E.Relationship.Multiplicity = M:1 AND
 E.Relationship.Has_Attribute = False
```

For each instance of entity “Relationship”, which has the attributes, or has the multiplicity “1:1” or “M:M”, it is necessary to create the single table that contains the key columns of each entity involved in relationship. We call this rule “Relationship_MM”, it has the following appearance:

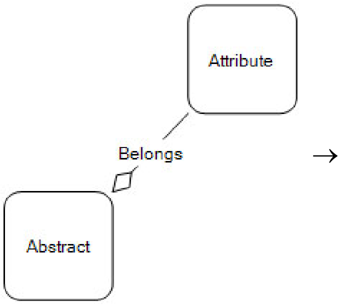


```
CREATE TABLE <<E.Relationship.Name>>
(id INTEGER primary key,
<<E.Relationship.LeftEntity>>ID INTEGER,
<<E.Relationship.RightEntity>>ID INTEGER)
ALTER TABLE <<E.Relationship.Name>> ADD
FOREIGN KEY
(<<E.Relationship.LeftEntity>>ID)
REFERENCES <<E.Relationship.LeftEntity>>
(id)
ALTER TABLE <<E.Relationship.Name>> ADD
FOREIGN KEY
(<<E.Relationship.RightEntity>>ID)
REFERENCES <<E.Relationship.RightEntity>>
(id)
```

This rule contains the constraint on the pattern occurrence:

```
E.Relationship.Multiplicity = M:M OR
E.Relationship.Multiplicity = 1:1 OR
 E.Relationship.Has_Attribute = True
```

The rule “Attribute” adds the columns corresponding to attributes of instances of entities and relationships to the created tables:



```
ALTER TABLE
<<E.Abstract.Name>>
ADD <<E.Attribute.Name>>
<<E.Attribute.Type>>
```

Let’s consider an example, apply the described transformation to the model “University” presented in Fig. 3.

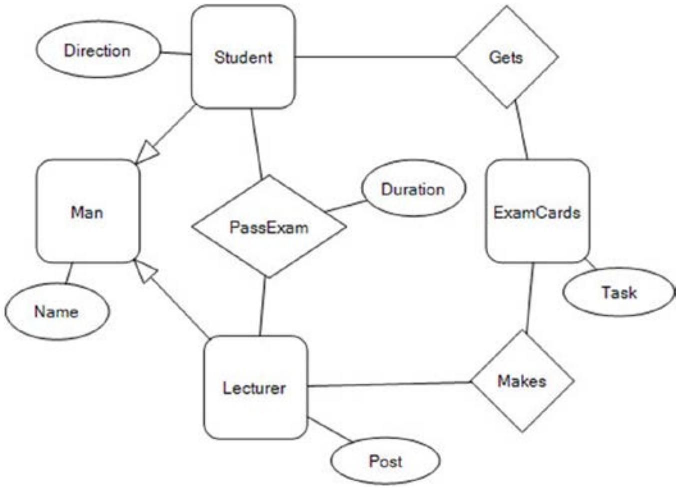


Fig. 3. Simplified model “University” on the ERD notation

As the result the following text has been generated by the MetaLanguage system:

```
CREATE TABLE Man (id INTEGER primary key)
CREATE TABLE Student (id INTEGER primary key)
CREATE TABLE Lecturer (id INTEGER primary key)
CREATE TABLE ExamCards (id INTEGER primary key)
ALTER TABLE Lecturer ADD ExamCardsID INTEGER
ALTER TABLE Lecturer ADD FOREIGN KEY
(ExamCardsID) REFERENCES ExamCards (id)
ALTER TABLE ExamCards ADD StudentID INTEGER
ALTER TABLE ExamCards ADD FOREIGN KEY (StudentID)
REFERENCES Student (id)
CREATE TABLE PassExam (id INTEGER primary key,
StudentID INTEGER, LecturerID INTEGER)
ALTER TABLE PassExam ADD FOREIGN KEY (StudentID)
REFERENCES Student (id)
ALTER TABLE PassExam ADD FOREIGN KEY (LecturerID)
REFERENCES Lecturer (id)
ALTER TABLE Student ADD ManID INTEGER
ALTER TABLE Student ADD FOREIGN KEY (ManID)
REFERENCES Man (id)
ALTER TABLE Lecturer ADD ManID INTEGER
ALTER TABLE Lecturer ADD FOREIGN KEY (ManID)
REFERENCES Man (id)
ALTER TABLE Man ADD Name nvarchar(MAX)
ALTER TABLE PassExam ADD Duration nvarchar(50)
ALTER TABLE Lecturer ADD Post nvarchar(50)
ALTER TABLE Student ADD Direction nvarchar(MAX)
```

It should be noted that this transformation does not take into account complex conversions the ERD notation to the database schema, for example, those which would allow to create single dictionary table on the base of attribute, because it requires a special description language of templates and it is one of the areas for further research. Although such conversion could be done by adding to the entity "Attribute" the attribute "Is_a_Dictionary" and setting the constraints on pattern occurrence.

Transformation "model-model". Transformation of this type allows to produce conversion of model from one notation to another or to perform any operations over model (creation of new elements, reduction, etc.). Such transformation will allow to export model to external systems, and to provide the ability to convert the domain-specific language that was created by the user in one of most common modeling language, for example, UML, ERD, IDEF0, etc.

The left-hand side of a production rule of this type transformation is some fragment of the source metamodel, and the right-hand side of the rule is some fragment of the target metamodel. At the production rule definition also it is necessary to describe the rules for converting the attributes of entities and relationships. The created model should not contain dangling pointers, therefore the process of the transformation executions begins with the creation of entity instances and only then instances of relationships are created. If in the process of model building the dangling pointers are still found the system will delete them.

At transformation execution it is necessary to consider the following elementary conversions:

- conversion "entity → entity";
- conversion "relationship → relationship";
- conversion "entity → relationship";
- conversion "relationship → entity".

Let's suppose that in the source model the instances of entities and relationships of pattern are already found.

For fulfillment of the conversion $ee: Ent_L \to Ent_R$ it is necessary to create in the new model the instance $EntI_R$ of the appropriate entity from a rule right-hand side and to perform transformation of attributes. The created instance of entity will have the same name, as the name of source entity instance.

For execution the conversion $rr: Rel_L \to Rel_R$ at first it is necessary to found in the source model the instances of entities $RelI_L.SEI$ and $RelI_L.TEI$, which are connected by the relationship instance $RelI_L$, then the images of these instances should be found in the new model, and an instance of the relationship from a rule right-hand side should be lead between them. After that it is necessary to fulfill transformation of attributes.

For fulfillment of the conversion $er: Ent_L \to Rel_R$ it is necessary to find in source model the nodes $EntI_S$, $EntI_T$ which are adjacent to entity instance $EntI_L$. Let's denote their images in the target model as *Source* and *Target*. In the target model the relationship instance $RelI_R$ between nodes *Source* and *Target* should be lead. Further it is necessary to execute defined transformation of attributes.

Conversion $re: Rel_L \to Ent_R$ transforms the instance of relationship $RelI_L$ found in the source model to the entity instance $EntI_R$ of target model. For conversion execution it is necessary to create the entity instance $EntI_R$, to perform the specified transformation rules of attributes. The name of $EntI_R$ will be the same as the name of the relationship instance $RelI_L$. At the next step it is necessary to find entities instances $RelI_L.SEI$, $RelI_L.TEI$, which are connected by relationship instance $RelI_L$.

Further the instances of relationships that connect an entity instance $EntI_R$ with nodes *Source* and *Target*, which are images of the nodes $RelI_L.SEI$ and $RelI_L.TEI$, accordingly, are created with keeping of orientation of relationship instance.

It is possible to present the rest conversions of "model-model" type by a combination of these elementary operations.

Let's consider an example, perform the transformation of the model on ERD notation to UML Class Diagrams.

Since the transformation is done at the metamodel level, then at the first step it is necessary to create/open source and target metamodels. The ERD metamodel was presented in the Fig. 2. Metamodel of UML Class Diagrams is shown in the Fig. 4. It contains the following elements: the entity "Class" and three relationships "Inheritance", "Association", "Aggregation". Let's define the production rules that determine the transformation.

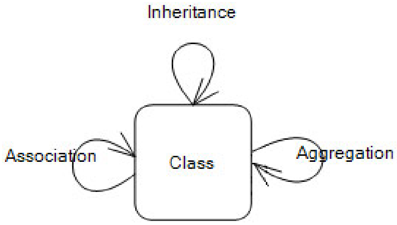
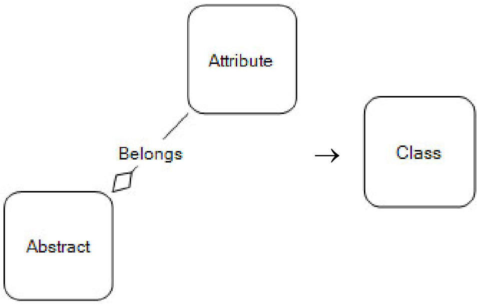


Fig. 4. Metamodel of UML Class Diagrams

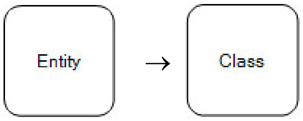
The rule "Abstract-Class" allows to convert the instances of entities "Entity" and "Relationship", which are connected at least with one instance of entity "Attribute", to the instance of entity "Class".

This rule has the following appearance:



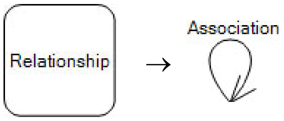
The rule "Entity-Class" allows to convert the instance of entity "Entity", which is not associated with any instance of the entity "Attribute", to the instance of an entity "Class".

The rule has the following form:

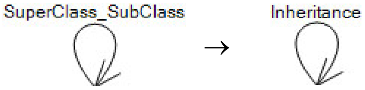


The rule "Relationship-Association" converts instances of the entity "Relationship" of the source model to instances of the relationship "Association" of the target model.

This rule looks like:



The rule "Inheritance" puts in correspondence to each instance of the relationship "SuperClass_SubClass" of source model a particular instance of the relationship "Inheritance" of target model. This rule has the following form:



After definition of all rules, which are included in the transformation, it is possible to execute conversion on a specific model. Let's perform this transformation on the considered earlier model "University" (see Fig. 3). The result of the transformation execution is presented in Fig. 5.

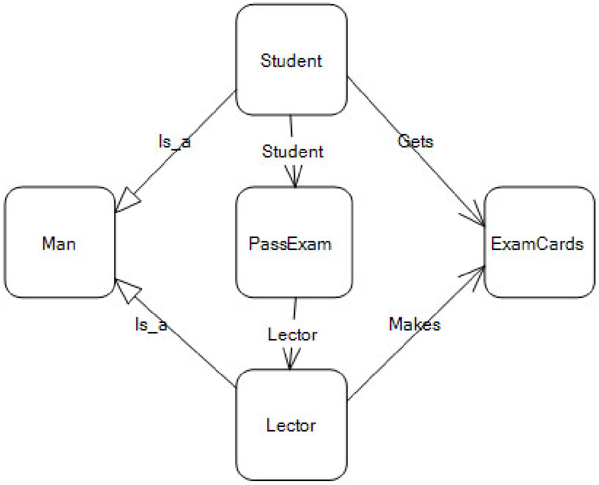


Fig. 5. Simplified model "University" in the Class Diagrams notation, generated by MetaLanguage system

Vertical transformation is a conversion of model, described at one level of hierarchy, to model presented at other level. Transformation of model allocated at higher level of hierarchy to model of subordinate level corresponds to operation of creation of model allocated at subordinate level. Inverse transformation allows to make interpretation of subordinate level model, to define types of its elements, to fulfil various operations over this model.

This mapping allows to support metamodels and created on their basis models in a consistent state. At metamodel modification the MetaLanguage system automatically makes all necessary changes in appropriate models.

Let's consider the process of vertical models transformations in more details.

If the model "University" is loaded in the MetaLanguage system as a metamodel, it will play the role of the domain-specific language and the models can be created on its basis. Let's construct on the basis of the domain-specific metamodel "University" the model "Exam". This model contains the following elements (see Fig. 6):

- "Test", "Essay" are instances of the entity "ExamCards";
- "Full-time student", "Extramural student" are instances of the entity "Student";
- "Professor", "Senior lecturer" are instances of the entity "Lecturer";
- "Name" is instance of the entity "Name";
- "Writes", "Solves" are instances of the relationship "Gets";
- "Checks", "Prepares" are instances of the relationship "Makes".

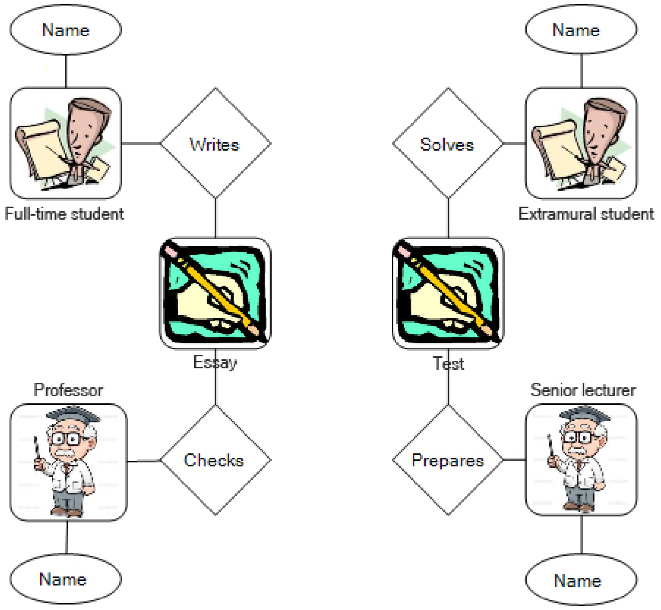


Fig. 6. Simplified model "Exam"

Thus, at creation of metamodel "ERD" the mapping of metalanguage constructions in metamodel entities and relationships is fulfilled. So the metalanguage construction "Entity" is mapped in the entities "Abstract", "Attribute", "Entity", "Relationship".

Then at construction of the domain-specific metamodel "University" the elements of metamodel "ERD" are mapped in instances of entities and relationships of the metamodel "University". For example, on the basis of entity "Attribute" its instances "Direction", "Duration", "Name", "Post", "Task" are built.

At the creation of model "Exam" the entities and relationships of the domain-specific metamodel "University" are mapped in elements of the model "Exam". So on the basis of entity "Lecturer" the elements "Professor", "Senior lecturer" are created.

At the stage of models validation and transformation the MetaLanguage system fulfills interpretation of models elements at various hierarchy levels. So at transformation of the domain-specific metamodel "University" in the SQL-query the language workbench should determine with what entities and relationships the elements of metamodel "University" are created, since transformation rules are described at level of metamodels. For example, at fulfillment of the previously described transformation "model-text" the MetaLanguage system will determine that elements "Man", "Student", "Lecturer", "ExamCards" are instances of the entity "Entity" and will apply to them the transformation rule "Entity".

V. Conclusion

Models transformations are a central part of the model-based approach to system development, because an existence in one system of models, which are fulfilled from the different points of view, with a different level of details and with use of different modeling languages, requires of existence of model transformation tools, which allow to convert models both between various levels of hierarchy, and within one level (at transition from one modeling language to another).

The presented approach has been implemented in a transformer of MetaLanguage system. This component allows to convert models described on visual domain-specific languages to text or other graphical models. The component has a convenient and simple user interface, therefore not only professional developers, but also domain specialists, for example business analysts, can work with it.

With the usage of this approach some languages and models have been developed. As example, the domain specific languages for the queuing system simulation have been designed and rules for transformation of visual simulation models into code in GPSS language have been described [26]. Generated model has been used for simulation running.

References

[1] A. Kleppe, J. Warmer, W. Bast, *MDA explained. The model-driven architecture: practice and promise*. Reading: Addison-Wesley, 2003, 170 p.

[2] S. Sendall, W. Kozaczynski, "Model transformation: the heart and soul of model-driven software development", *IEEE Software*, vol. 20, pp. 42–45, 2003.

[3] A. O. Sukhov, "The language workbench for visual domain-specific modeling languages creation", *Fundamental Researches*, vol. 4, pp. 848-852, 2013.

[4] E. Grabska, B. Strug, "Applying cooperating distributed graph grammars in computer aided design", *Parallel Processing and Applied Mathematics*, vol. 3911/2006, pp. 567–574, 2006.

[5] T. Mens, K. Czarnecki, P. V. Gorp, "A taxonomy of model transformations", *Electronic Notes in Theoretical Computer Science*, vol. 152, pp. 125–142, 2006.

[6] U. Montanari, F. Rossi, "Graph rewriting, constraint solving and tiles for coordinating distributed systems", *Applied Categorical Structures*, pp. 333–370, 1999.

[7] B. Konig, "Analysis and verification of systems with dynamically evolving structure", habilitation thesis [Online]. Available: http://jordan.inf.uni-due.de/publications/koenig/habilschrift.pdf.

[8] J. Rekers, A. Schuerr "A graph grammar approach to graphical parsing", in *Proc. of the 11th IEEE International Symposium*, Washington, 1995, pp. 195–202.

[9] H. Ehrig, K. Ehrig, U. Prange, G. Taentzer, *Fundamentals of algebraic graph transformation*. New York: Springer-Verlag, 2006, 388 p.

[10] A. Corradini, U. Montanari, F. Rossi, H. Ehrig, R. Heckel, M. Loewe, "Algebraic approaches to graph transformation. Part I: basic concepts and double pushout approach", *Handbook of Graph Grammars and Computing by Graph transformation*, vol. 1, pp. 163–246, 1997.

[11] H. Ehrig, R. Heckel, M. Korff, M. Loewe, L. Ribeiro, A. Wagner, A. Corradini, "Algebraic approaches to graph transformation. Part II: single pushout approach and comparison with double pushout approach", *Handbook of Graph Grammars and Computing by Graph Transformation*, vol. 1, pp. 247–312, 1997.

[12] A. Agrawal, G. Karsai, S. Neema, F. Shi, A. Vizhanyo, "The design of a language for model transformations", *Journal on Software and Systems Modeling*, vol. 5, pp. 261–288, 2006.

[13] D. Balasubramanian, A. Narayanan, C. P. Buskirk, G. Karsai, "The graph rewriting and transformation language: GReAT", *Electronic Communications of the EASST*, vol. 1, pp. 1–8, 2006.

[14] T. Gardner, C. Griffin, J. Koehler, R. Hauser, "A review of OMG MOF 2.0 Query/Views/Transformations submissions and recommendations towards the final standard", in *Proc. of the 1st International Workshop on Metamodeling for MDA*, York, 2003, pp. 1–20.

[15] P. Stevens, "Bidirectional model transformations in QVT: semantic issues and open questions", *Model Driven Engineering Languages and Systems*, vol. 4735/2007, pp. 1–15, 2007.

[16] G. Csertan, G. Huszerl, I. Majzik, Z. Pap, A. Pataricza, D. Varro, "VIATRA – visual automated transformations for formal verification and validation of UML models", in *Proc. of the 17th IEEE International Conference on Automated Software Engineering*, Washington, 2002, pp. 267–270.

[17] A. Balogh, D. Varro, "Advanced model transformation language constructs in the VIATRA2 framework", in *Proc. of the ACM Symposium on Applied Computing*, New York, 2006, pp. 1280–1287.

[18] V. Chiprianov, Y. Kermarrec, P. D. Alff, "An approach for constructing a domain definition metamodel with ATL", in *Proc. of the 1st International Workshop on Model Transformation with ATL*, Nantes, 2009, pp. 18–33.

[19] F. Jouault, F. Allilaire, J. Bezivin, I. Kurtev, "ATL: a model transformation tool", *Science of Computer Programming*, vol. 72, pp. 31–39, 2008.

[20] D. Varro, Z. Balogh, "Automating model transformation by example using inductive logic programming", in *Proc. of the ACM Symposium on Applied Computing*, New York, 2007, pp. 978–984.

[21] M. Wimmer, M. Strommer, H. Kargl, G. Kramler, "Towards model transformation generation by-example", in *Proc. of the 40th Annual Hawaii International Conference on System Sciences*, Washington, 2007, pp. 1–10.

[22] J. R. Ullmann, "An algorithm for subgraph isomorphism", *Journal of the Association for Computing Machinery*, no. 23, pp. 31–42, 1976.

[23] D. Schmidt, L. Druffel, "A fast backtracking algorithm to test directed graphs for isomorphism using distance matrices", *Journal of the Association for Computing Machinery*, no. 23, pp. 433–445, 1976.

[24] L. P. Cordella, P. Foggia, C. Sansone, M. Vento, "An improved algorithm for matching large graphs", in *Proc. of the 3rd Workshop on Graphbased Representations in Pattern Recognition*, Ischia, 2001, pp. 149–159.

[25] B. D. McKay, "Practical graph isomorphism", *Congressus Numerantium*, vol. 30, pp. 45–87, 1981.

[26] E. B. Zamyatina, L. N. Lyadova, A. O. Sukhov, "An approach to integration of modeling systems and information systems on the basis of DSM-platform MetaLanguage", in *Proc. of the 4th International Conference Information Systems Development Technologies*, Gelendzhik, 2013, pp. 61–70.

Future Satellite Systems for Emergency Communications Situations

Haitham Cruickshank
ETSI SatEC group, STF 472 team
h.cruickshank@surrey.ac.uk

Anton Donner
ETSI SatEC group, STF 472 team
Anton.Donner@dlr.de

Robert Mort
ETSI SatEC group, STF 472 team
mort.robert@gmail.com

Egil Bovim
ETSI SatEC group, STF 472 team
egil.bovim@gmail.com

Julian Sesena
ETSI SatEC group, STF 472 team
julian.sesena@wirelesspartners.es

Abstract— **For many applications, Satellite is the preferred delivery mechanism due to its wide area coverage and multicasting capacity. Major emergencies or disasters may result in a need for additional satellite resources in local telecommunications networks, especially if they are damaged or overloaded, in order to maintain or enhance the ability of rescue workers to respond and coordinate their activities effectively. This paper presents the current ongoing work within the ETSI SatEC working group on small and large scale emergency scenarios. An overview of user requirement is also presented together with communication network requirements and flows. Future work will focus on producing a more detailed satellite network requirements and a network topological model showing how end-users satellite terminals are deployed/move on their activity field.**

Keywords— satellite networks; Earthquakes, Mass casuality incidents, communication requirements

I. Introduction

Major emergencies or disasters may result in a need for additional resources in local telecommunications networks, especially if they are damaged or overloaded, in order to maintain or enhance the ability of rescue workers to respond and coordinate their activities effectively [4], [5]. Satellites can play a role in replacing or supplementing other telecommunications links in these scenarios. For example satellite systems can provide:

- Broadband and secure communication facilities anywhere/anytime in locations where no other facilities are available.
- Temporary replacement of broken/saturated infrastructures by means of backhauling.
- Fast deployment of temporary communication networks in emergency situations.

Hence a set of requirements for such links needs to be established in emergency situations.

This work within ETSI is also a response to EC mandate M/496, specifically dossier 9 "Disaster Management" part 2: “Emergency Telecommunication Services” which aims to support standardization for the optimal needs of the emergency responders.

The focus of current work is two types of emergency scenarios [6], [7]: Major EarthQuake (EQ) in an urban environment and public Mass Casualty Incident (MCI) in the countryside. These scenarios are chosen because the lack of telecommunication infrastructure highlights the role of satellites that can play a vital role in quickly establishing the needed communication networks.

II. Mass casuality incident (MCI) scenario

The MCI scenario is defined for the evaluation and dimensioning of small scale satellite-based emergency telecommunications. This scenario includes potential roles for satellite systems for the telecommunication services identified. The aim is to define firstly a disaster scenario, then the general communication needs of the actors involved. The future objective is the more detailed results of topology modelling of these communications requirements are provided.

Major incidents or accidents such as mass transportation accidents can lead to MCIs. They are characterized by a misbalance between available Emergency Medical Services (EMS) resources and patients requiring medical care. The threshold for an MCI depends on deployable resources and will typically differ between e.g., urban and rural settings. The rural setting is the focus of this work because there might a lack of communications facilities at the MCI site.

The main objective of MCI response is to provide fast and adequate help to a maximal number of patients. This requires a comprehensive situation overview. MCIs are typically sudden incidents with little lead time and stringent requirements in terms of logistics efficiency. E.g., the out-of-hospital time for trauma patients should be less than one hour (“golden hour of shock”).

In order to meet these requirements regular EMS procedures have to be switched to a temporary “overload mode” involving additional operational and tactical structures. Most of the relevant decisions have to be taken on-site, but a major obstacle is that the situation continuously evolves over time with arriving rescue forces, transport means, and updated incident information. Additionally, the MCI can have a spatial dimension complicating the situation assessment and information exchange.

An example scenario is a mass transportation accident in a rural environment and covers a small geographic area. MCIs may also occur in wide-area incidents. Likewise causes of MCIs are of various natures and not restricted to man-made disasters. Examples are epidemics and natural disasters. As such, an MCI is a consequence of many persons requiring help in comparison to limitations in:

- Rescue personnel in general and more specifically medical personnel like paramedics and emergency physicians.
- Medical supplies like dressing material, infusions, and pharmaceuticals.
- Medical tools like handbarrows, suction pumps, and rebreathing devices.
- General shelter equipment like blankets, and water.
- Transport means like ambulance cars, busses, or helicopters.
- Hospital treatment capacity and shelter capacity.

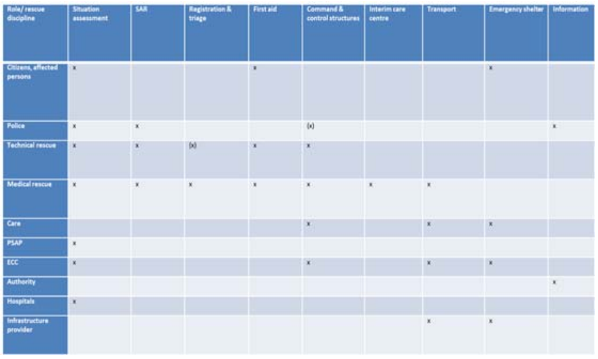
A. *MCI disaster response actions*

Efficient MCI handling requires several concurrent actions. All affected persons have to be assessed first (“triage”) so that their further treatment can be prioritized according to their actual urgency and available resources in terms of medical personnel, transport means and hospital capacities. In other words: each patient should get the right treatment at the right time. Here is a brief overview of the most important tasks:

- Situation/resource assessment: such as reports by affected witnesses or by indirectly affected citizens.
- Search And Rescue (SAR): such as rescue/evacuation out of hazard zone (e.g., firefighting, breathing protection and surface water rescue).
- Command and control structures: such as set-up of Emergency and Field Emergency Control Centres (FECC) for involved disciplines (such as medical, fire and police entities)
- Information and communication to the public: setup Public Safety Answering Point (PSAP).

Table I provides an overview of MCI entities and their possible tasks.

TABLE I. ENTITIES AND TASKS DURING MCI RESPONSE



Role/ rescue discipline	Situation assessment	SAR	Registration & triage	First aid	Command & control structures	Interim care centre	Transport	Emergency shelter	Information
Citizens, affected persons	x			x				x	
Police	x	x			(x)				x
Technical rescue	x	x	(x)	x	x				
Medical rescue	x	x	x	x	x	x	x		
Care					x		x	x	
PSAP	x								
ECC	x				x		x	x	
Authority									x
Hospitals	x								
Infrastructure provider							x	x	

In Fig. 1, a typical MCI process chain with the patient flow is shown. This diagram is a simplification of actual rescue operation dynamics since it does not consider the spatial and temporal dimension of the incident. The color code used in the picture representing patient exigency (urgent demand) and the triage categories themselves differ between countries.

Ideally, the triage areas should be optimally placed in relation to the logistics, but they are self-organizing. The incident commander decides if an interim care center has to be set up. Authorities and task forces are involved in major MCIs or if the MCI is a part of a bigger disaster.

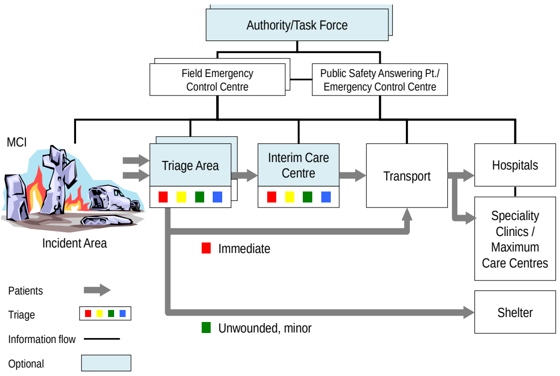


Fig. 1. Theoretical MCI process chain

III. EARTHEQUAKE (EQ) SCENARIO

The EQ defines a reference scenario for an earthquake which is relevant for the evaluation and dimensioning of large scale satellite-based emergency telecommunications. This scenario includes potential roles for larger satellite systems for the telecommunication services identified. The services defined for these scenarios are limited to safety services (i.e. not security such as law enforcement). The aim is to define firstly a disaster scenario, then the general communication needs of the actors involved. The future objective is the more detailed results of topology modelling of these communications requirements are provided.

An earthquake scenario is defined in terms of its main constituent events and secondly by its physical consequences. Thus the response actions by emergency forces to this scenario are defined in terms of the casualties involved, the actors and organizations, overall operations modes, duration and dimensioning factors etc.

This scenario is chosen to be sufficiently generic to be considered representative of many potential future earthquakes, and thus to allow relevant communication characteristics for current and future needs to be established.

The main assumption is that the earthquake affects an urban area. The Earthquake is assumed to be of a magnitude sufficient to cause a multitude of physical effects, such as collapsed buildings, flash-floods/tsunamis, disruption of infrastructure with resulting traffic accidents, lack of power, lack of telecommunications, fires, risks of chemical accidents etc. Each of these incidents may not differ much from isolated similar incidents of this nature, but the added challenge is that the incidents happen at the same time, thus reinforcing the effects and strains on available resources. Here are some example effects of an earthquake:

- Physical Effects such as collapse of buildings, fire and chemical accidents.
- Disruption infrastructure such as power, water, sanitation and transport and telecommunications.
- Disruption to services: such as emergency services (police, fire and health services).

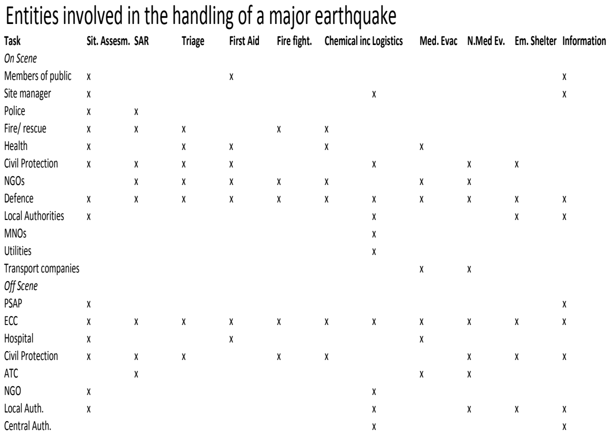
A. *EQ disaster response actions*

In addition to the task describe in the MCI section above (see II.A), the following extra tasks are required due to the large scale of EQ compared to MCI scenario:

- Extra situation/resource assessment for dangers such as flooding and chemical accidents.
- Evacuation of population: such as evacuation to temporary dwellings: Emergency services/ Rescue Personnel/ Local Authorities
- Emergency shelters: such as temporary housing, food supply, water supply, electricity, sewage: Local Authorities/ Civil Protection.

Depending on local/ national organization of services and division of tasks/ responsibilities, the entities involved and their individual areas of work may differ. Table below provides an overview of entities that are most commonly involved, and for which responses they are mainly involved. Depending on the severity of the earthquake, resources may be drawn nationally, regionally and internationally (e.g. involving fire-fighters from several countries). Table II provides an overview of EQ entities and their possible tasks.

TABLE II. ENTITIES AND TASKS DURING EQ RESPONSE



Entities involved in the handling of a major earthquake

Task	Sit. Assesm.	SAR	Triage	First Aid	Fire fight.	Chemical inc	Logistics	Med. Evac	N.Med Ev.	Em. Shelter	Information
On Scene											
Members of public	x			x							x
Site manager	x						x				x
Police	x	x									
Fire/ rescue	x	x	x		x	x					
Health	x		x	x		x		x			
Civil Protection	x	x	x	x			x		x	x	
NGOs		x	x	x	x	x		x	x		
Defence	x	x	x	x	x	x	x	x	x	x	x
Local Authorities	x						x			x	x
MNOs							x				
Utilities							x				
Transport companies								x	x		
Off Scene											
PSAP	x										x
ECC	x	x	x	x	x	x	x	x	x	x	x
Hospital	x			x				x			
Civil Protection	x	x	x		x	x			x	x	x
ATC		x						x	x		
NGO	x						x				
Local Auth.	x						x		x	x	x
Central Auth.							x				x

IV. COMMUNICATION LINKS BETWEEN EMERGENCY ENTITIES

Fig. 1 depicts the main involved entities/roles and typical communication links [1], [2], [3]. The involved disciplines such as technical/medical rescue, care and police are shown as “rescue disciplines A, B, C…”. Depending on the incident there might be none, one or many FECCs for each rescue discipline operating in a hierarchical structure. In case of an MCI the medical rescue command structure might have dedicated FECCs for the triage area, for the interim care centre, and for the transport section. For the EQ scenario, there will be multi-disciplinary rescue teams: for example discipline A (Fire& Rescue team), Discipline B (Health and medical emergency team) and Discipline C (police or state armed forces).

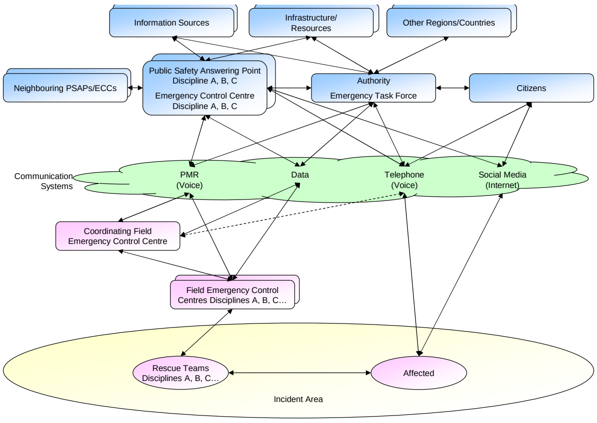

Information Sources
Infrastructure/ Resources
Other Regions/Countries
Neighbouring PSAPs/ECCs
Public Safety Answering Point Discipline A, B, C
Emergency Control Centre Discipline A, B, C
Authority
Emergency Task Force
Citizens
Communication Systems
PMR (Voice)
Data
Telephone (Voice)
Social Media (Internet)
Coordinating Field Emergency Control Centre
Field Emergency Control Centres Disciplines A, B, C…
Rescue Teams Disciplines A, B, C…
Affected
Incident Area

Fig. 2. typical communication links between involved emergency entities/roles

Some the communication flows between various entities (in Fig. 2) are described below:

- FRs – FRs: For example, the First Responders (FRs) in each emergency team (Discipline A, B or C) has its own and unique talk group.
- FRs – corresponding ECC/FEEC: Bidirectional voice communication is allowed between the FRs and their command posts since they share the same

voice group. Bidirectional data communication is allowed also between FRs and their ECC/FEEC primarily for geo-positioning applications. Special data services are required for the Fire & Rescue discipline in hot area since they may carry special equipment which needs to transmit data to the corresponding ECC/FEEC. Examples of such equipment are wearable biometric sensors, radiation or gas sensors, video cameras, positioning equipment. The Command Posts (CP) should have equipment to receive, integrate and display all these data acquiring valuable risk management information. Crucial actions are allowed, i.e. FRs could be rescued if irregular vital signs are observed. FRs will receive alerts, positioning information, commands to proceed, etc, through voice communication or data displayed in their user interface.

- ECC – ECC: In the case several ECCs are deployed, they all should have bidirectional communication among them sharing valuable risk management information from their corresponding Coordinating ECC or Emergency Agency.

- ECC – EATF: All ECCs deployed in the disaster area and the global coordination centre (Emergency Authority Task Force, EATF) in Figure 1, should have bidirectional voice and data communication among them sharing valuable risk management information or commands to proceed.

- ECCs – External entities: Certain ECCs have bidirectional communications with external entities with or sometimes without the intervention of the EATF, such as Fire & Rescue communicate with medical vehicles with hospitals. In addition to current operational practices where just voice calls are performed, paramedical users have suggested to enable ambulances to send data files with injured victims lists or electrocardiogram traces to the hospital to reduce reaction time attending the victims.

- ECCs – External information acquisition: It is desirable for CPs to have external connections in order to enhance effectiveness of the operational procedures. One example is access to detailed maps for geo-localization support units would receive detailed information of the target location. Another application could be accessing the Fire Risk Assessment Indices of the affected area which takes into account baseline vegetation information (vigour, condition, etc). Access near real time satellite images is a new service provided by current observation LEO satellites in charge of image acquisition of the hot spots within several hours from the start of the emergency.

V. Communication Service Requirements

With respect to Fig. 2, the efficient exchange of information within a single discipline team, between ECCs, between a temporary task force/command centre and a permanent PSAP/ECC may be facilitated by a number of communication services, described below [2], [8]:

A. *Speech services*

Speech services are currently the most instinctive and most used communication services in emergencies. As such there exist several universal requirements, characterized by:

- Speech intelligibility and quality: that received speech can be understood reliably and in some cases high speech quality is desirable.
- Call setup-time: short call set-up times enable rapid communication of relevant information.

- End to end delay: in addition to the call set-up delay, it is recognized that where a duplex voice communication system that imposes an end to end delay of over 500 ms, there is degradation in the voice quality (ITU T Recommendation G.114).

Also underlying networks (e.g. satellite and terrestrial) should have the capability to handle prioritized calls correctly, including the capability of pre-emption of un-prioritized calls. Transit networks should convey priority related signalling in order to support end-to-end priority. The following list describes some of the required speech services.

Point To Point Speech Services: Point to point duplex voice communications are required for many instances to provide communications, particularly between different authorities e.g. between commanders of different emergency services, between emergency service staff and external specialists.

Group Speech Services: Two examples of groups are:

- Talk group: Point-to-multipoint group addressed communication established within a selectable predefined area.

- Emergency services call (authority to authority)

Push-To-Talk (PoC)/Command and Control (C&C) features: PoC helps to avoid network congestion by transmitting voice over a data channel (GPRS, UMTS) and thus can be used even in times of high traffic on the communication network.

B. *Video Tele-Conferencing (VTC)*

VTC may be required to enable effective coordination between services at a command level or below. VTC services may be utilized to provide reconnaissance information from the incident back to control rooms. Note that near-real-time video streaming can be considered as a data service (see below).

C. *Data services*

Data services are used to provide a large number of applications which can have widely differing requirements in terms of capacity, timeliness and robustness of the data service. Ideally, the communication networks should support the required data throughput and minimize end to end delay, especially for applications such as real time video streaming.

Noting the extreme circumstances which may be in force during an emergency, it may be desirable for networks to degrade gracefully when user requirements exceed the agreed levels of service.

Table III shows the diverse needs of data applications. Where data applications share the use of a data transmission capability, provision of sufficient capacity and effective management must be provided to ensure application data is communicated appropriately. The definition of Table 3 categories are:

- **Throughput:** data volume in a given time.
- **Timeliness:** importance of the information arriving within an agreed timeframe.
- **Preservation of data integrity:** how (reliable) free from bit errors the information transmission needs to be. E.g. a bitmap image with some errors is still useable; a jpg image with some bit errors may be unreadable.

TABLE III. REQUIREMENTS ON DATA APPLICATIONS [2]

Service	**Throughput**	**Timeliness**	**Need for preservation of data integrity**
Email	Medium	Low	Low
Imaging	High	Low	Variable
Digital mapping/ Geographical information services	High	Variable	Variable
Location services	Low	High	High
Video (real time)	High	High	Low
Video (slow scan)	Medium	Low	Low
Data base access (remote)	Variable	Variable	High
Data base replication	High	Low	High
Personnel monitoring	Low	High	High

D. *Paging (short message) Services*

Paging services are used by a variety of authorities in order to contact their personnel, and paging services are available from a variety of networks and technologies. The network needs to be able to identify the requested emergency agent(s), and then deploy the appropriate technology to contact them. This requirement may encompass different communication network technologies, services and applications such as paging, presence, texting, etc.

E. *Status Monitoring and Location Services*

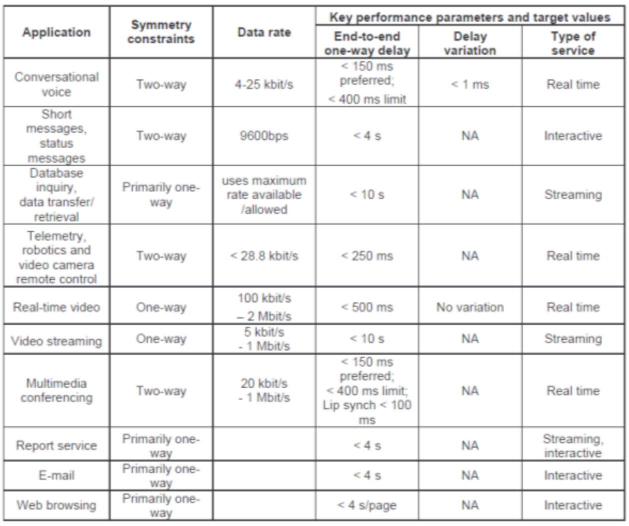
Status monitoring includes a wide variety of parameters, e.g. breathing air tank levels, accountability monitoring, distress buttons and vital signs monitoring. Location services provide real-time information regarding the position of personnel or vehicles to a Command Post (CP). This information may also include status information regarding the person or vehicle. The service may require frequent transmissions to update position; the amount of data transmitted is likely to be small when location is based on satellite-based solutions, but can be quite extensive when location is to be calculated inside buildings as other technologies may have to be used. Location reporting services may be one-way with no acknowledgement, necessitating a robust communication mechanism. Position information may be considered sensitive in some emergencies and may require security mechanisms to protect the data.

Based on an assessment of the currently used communication services within the public safety sector it can be concluded that public safety mobile communications are voice-based with widespread use of group calls (“network-centric”), also called talkgroups. These are called Push-to-talk calls (PTT calls) Point-to-point voice calls (P2P calls) are also used specially by emergency managers. According to user comments, positioning information of emergency units deployed in the disaster area is also currently in used in most of Public Agencies that counts on narrowband data channels of TETRA technology. However, there is a trend towards using a range of data applications alongside traditional voice applications to enhance communications. Data services have widely differing requirements in terms of capacity, timeliness and robustness of the data service.

Table below shows a summary of characteristics of the identified communication service. Satellite and terrestrial communication networks should be able to provide the required data rates and QoS provisioning parameters.

It should be noted the this work is still at an early stage where more detailed mapping of the scenario parameters and communication service requirement will be contacted a later stage together with satellite network topology modelling in these scenarios.

TABLE IV. CHARACTERISTICS OF THE IDENTIFIED COMMUNICATION SERVICES [9]



Application	Symmetry constraints	Data rate	Key performance parameters and target values		
			End-to-end one-way delay	Delay variation	Type of service
Conversational voice	Two-way	4-25 kbit/s	< 150 ms preferred; < 400 ms limit	< 1 ms	Real time
Short messages, status messages	Two-way	9600bps	< 4 s	NA	Interactive
Database inquiry, data transfer/ retrieval	Primarily one-way	uses maximum rate available /allowed	< 10 s	NA	Streaming
Telemetry, robotics and video camera remote control	Two-way	< 28.8 kbit/s	< 250 ms	NA	Real time
Real-time video	One-way	100 kbit/s – 2 Mbit/s	< 500 ms	No variation	Real time
Video streaming	One-way	5 kbit/s - 1 Mbit/s	< 10 s	NA	Streaming

Multimedia conferencing	Two-way	20 kbit/s
		- 1 Mbit/s
			< 150 ms preferred;
			< 400 ms limit;
			Lip synch < 100 ms
				NA	Real time

Report service	Primarily one-way		< 4 s	NA	Streaming, interactive
E-mail	Primarily one-way		< 4 s	NA	Interactive
Web browsing	Primarily one-way		< 4 s/page	NA	Interactive

VI. CONCLUSION

Major emergencies or disasters may result in a need for additional resources in local telecommunications networks, especially if they are damaged or overloaded, in order to maintain or enhance the ability of rescue workers to respond and coordinate their activities effectively. Satellites can play a role in replacing or supplementing other telecommunications links in these scenarios.

This paper has presented the current ongoing work within the ETSI SatEC working group on two types of emergency scenarios: Major earthquake in an urban environment and Mass casualty accident in the countryside. In both scenarios satellites can play a vital role in quickly establishing the needed communication networks. One future vision is t help with the private cloud to cash maps and frequently needed information plus synchronization of data between various ECCs, EATFS and some external entities.

A generalized overview of user requirement has been presented together with communication network requirements and flows. However, collecting the user requirements is still ongoing task. This will help to produce more detailed satellite network requirements and a textual, graphical or mathematical topological model showing how end-user communication equipment are deployed/move on their activity field.

REFERENCES

[1] ETSI TR 102 180: "Emergency Communications (EMTEL); Basis of requirements for communication of individuals with authorities/organizations in case of distress (Emergency call handling)".

[2] ETSI TS 102 181: "Emergency Communications (EMTEL); Requirements for communication between authorities/organizations during emergencies".

[3] ETSI TS 102 182: “Emergency Communications (EMTEL); Requirements for cimmunications from authorities/organizations to individuals, groups or the general public during emergencies”.

[4] ETSI TR 102 641: “Satellite Earth Stations and Systems (SES); Overview of present satellite emergency communications resources”.

[5] ETSI TR 103 166: "Satellite Earth Stations and Systems (SES); Satellite Emergency Communications (SatEC); Emergency Communication Cell over Satellite (ECCS)".

[6] ETSI TS 103 260-1: “Satellite Earth Stations and Systems (SES); Reference scenario for the deployment of emergency communications; Part 2: Earthquake”

[7] ETSI TS 103 260-2: “Satellite Earth Stations and Systems (SES); Reference scenario for the deployment of emergency communications; Part 2: Mass casualty incident in public transportation”

[8] ITU-T Recommendation G.1010: End-user multimedia QoS categories

[9] Widens project, "System requiremenst and first system architecture design". Deliverable D2.1. April 2004

Personalization of student in course management systems on the basis using method of data mining

Martin Magdin, Milan Turčáni

Abstract**—**Individualization of learning through ICT allows to students not only the possibility choose the time and place to study, but especially pace adoption of new knowledge on the basis of preferred learning styles. Analysis of learning processes should give the answer to difficult questions from pedagogical and psychological theory and practice. Count of scientific studies that should represent the results of systematic and long-term oriented studies in this area is still few. With the Learning styles and possibilities of their application in the context of e-learning addresses many experts. These experts predict that student should know, which Learning style is best for him, or predict that alone student knows when is the right time to try it differently. For determination of learning styles and personalization of student in Course Management System are used various techniques. In the paper we present the use non standard of techniques of data mining - data mining based on the use of interactive animations in e-learning courses. With this method of data mining we can get a complete overview of the activities of the student and on the basis of the definition of so-called social rules we know adjust to educational content.

Keywords**—**Personalization, Interactive animations, Data mining, Adaptive education, Simulation, Adaptive study support.

I. Introduction

MASS education in a classroom or with the help of classic e-learning is not able to respond to individual needs of a studying individual. Some students are restrained and bored by it, for some, on the other hand, it is too quick and they do not manage understand everything or the education style of each teacher does not have to be suitable for them [2]. Other students are satisfied with the pace of education, but they may not be satisfied with the teaching style of a particular teacher. Therefore, such students come to dislike the teachers and subjects they teach, which results in them having worse results [4, 8]. The suggested reasons lead to the idea of the optimization of the learning process through the use of individualization of education. Individualization of education represents each student's way of learning with regard to their previous knowledge, skills and their learning style [19, 16]. A set of attitudes and behaviors which determine an individual's preferred way of learning is considered as a learning style [14]. Learning styles have been a subject of extensive research [10, 11, 31], however the research focuses predominantly on their identification and classification. A team of experts dealt with the research and processing of the theory of learning styles in the Czech Republic [40, 36]. In Slovakia there exist only a few experts dealing with this sphere of pedagogy [5, 41]. The work by [23] served as a basis for the research of the Czech and Slovak experts.

In general, the purpose of the theory of processing and evaluation of individual styles of teaching is the proposal and solution of the problem of individualization of educational process. If we put together the essence and principle of e-learning and the request for personalized learning, we gain a relative new research area – adaptive learning. Optimal adaptive process will respect students' differences based on their learning style and with regard to their changing knowledge and skills during the course of the study in the course. On the basis of identification of personal characteristics and qualities, the students will be provided with a study material that suits them the most [20]. We assume that personally tailored education accenting student's requirements, preferences, and positive sides of learning (we do not support surface learning, remembering without understanding, etc.) will become an optimal and effective form of education. It will make new knowledge easier to remember and more permanent.

II. Adaptive educational theory (learning styles in e-learning)

Quantum of pedagogical and didactical principles (rules) form the theoretical basis for the formulation of adaptive educational theory (AET). These rules are based first of all on the following approaches:

- Komensky – systematic and methodical approach,
- Gagne – result can be accomplished by elementary steps,
- Bloom´s taxonomy – 6 levels of knowledge (remembering, understanding, applying, analyzing, synthetizing and the ability to evaluate information) for successful realization of partial obligations, which are intensified during the course of education as to their degree of difficulty,

Martin Magdin is with the Department of Informatics, Faculty of Natural Sciences, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 940 11 Nitra, Slovakia (phone: +421376408676; e-mail: mmagdin@ukf.sk).

Milan Turčáni is with the Department of Informatics, Faculty of Natural Sciences, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 940 11 Nitra, Slovakia (phone: +421376408671; e-mail: mturcani@ukf.sk).

- Theory of program teaching (Skinner) – division of the contents of education into smaller wholes, their verification and reaction to the comprehension of the contents of education,
- Adaptive hypermedia systems (Brusilovský) – feedback and evaluation of behavior of the student during the process of education – journaling the process of teaching.

By reason of inconsistently processed classification of learning styles it is possible to meet with various models of classification of learning styles in the area of AET (most frequently Shulman´s or Felder-Silverman´s model).

Shulman's model (TPCK) – Conjunction of the pedagogical and contentual dimension means understanding and solving the particular pedagogical situation with the use of suitable learning methods and forms with the aim to accomplish effectiveness of the educational process.

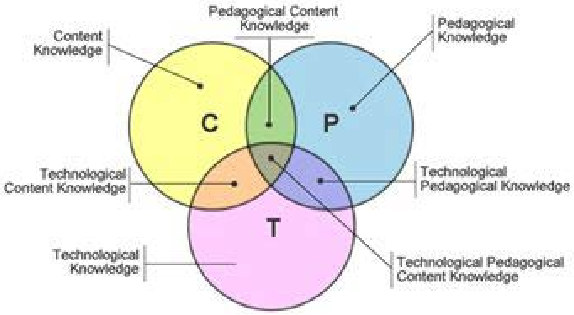

Pedagogical Content Knowledge
Content Knowledge
Pedagogical Knowledge
C
P
Technological Content Knowledge
Technological Pedagogical Knowledge
T
Technological Knowledge
Technological Pedagogical Content Knowledge

Fig. 1 Shulman's model TPCK [25]

Based on the initial version (1986), Mishra and Koehler gradually extended it by a new dimension – technological aspect in 2006. Conjunction of all 3 dimensions is the defined work of the teacher with the current ICT with the aim to optimize and increase effectiveness of the educational process.

Felder-Silverman' model (FSLSM) - Felder and Silverman (1988) advocate that students learn in different ways: by hearing and seeing; by reflecting and acting; reasoning either logically or intuitively; by memorizing and visualizing and drawing analogies; and, either steadily or in small bits and large pieces. They also advocate that teaching styles vary, such as an educator's preference for lecturing or demonstrating, or for focusing on principles or applications.

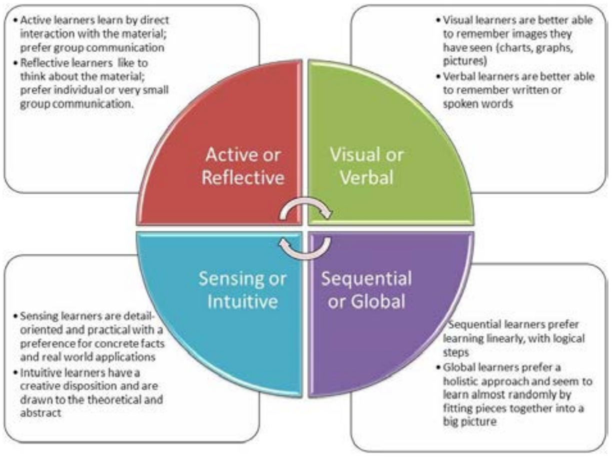

• Active learners learn by direct interaction with the material; prefer group communication
• Reflective learners like to think about the material; prefer individual or very small group communication.
• Visual learners are better able to remember images they have seen (charts, graphs, pictures)
• Verbal learners are better able to remember written or spoken words
Active or Reflective
Visual or Verbal
Sensing or Intuitive
Sequential or Global
• Sensing learners are detail-oriented and practical with a preference for concrete facts and real world applications
• Intuitive learners have a creative disposition and are drawn to the theoretical and abstract
Sequential learners prefer learning linearly, with logical steps
• Global learners prefer a holistic approach and seem to learn almost randomly by fitting pieces together into a big picture

Fig. 2 Felder-Silverman's model [6]

According to [28] for assess students' learning style according to FSLSM, the Index of Learning Styles (ILS) is generally used [13]. It contains 44 two choice questions distributed along the four learning style dimensions, where one choice increments and the other decrements the score of the particular dimension. The resulting index of preference for each dimension is expressed by an odd integer, ranging [-11, +11] since 11 questions are posed for each dimensions. The ILS questionnaire provides a very precise quantitative estimation of a learner's preference for each dimension of FSLSM.

III. Media elements and methods of data mining

Learning management systems are commonly used in e-learning, but provide low level of adaptivity. By combining adaptation and personalization into LMS, a new kind of tailored learning environments which motivate learners can be built [38].

In case of utilizing e-learning systems, it is inevitable to utilize various techniques of data mining in order to expressly define fruitfulness of the continuous study (increment of knowledge, skills and experiences of the students) and based on the results to design a suitable learning style for the student.

Moodle accumulate amount of information which is very valuable for analyzing students' [26]. For example they can record student activities, academic results, user's interaction data, etc. Although some platforms offer different reporting tools, do not provide however specific tools which allow educators to thoroughly track and assess all the activities performed by their learners and to evaluate the structure and contents of the course and its effectiveness in the learning process. Very promising area for attaining this objective is the use of data mining.

Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from the educational context [33]. Methods of EDM is the automatic extraction of implicit and interesting patterns from large data collections. This methods is mainly used in the last decade for improve e-learning systems [32]. Data mining can be applied to explore, visualize and analyze e-learning data in order to identify useful patterns, to evaluate web activity to get more objective feedback for teachers' instruction and to find out more about how the students learn.

According to [32] EDM is an iterative cycle which consists of the same four steps in the general data mining process as follows:

1. Collection data,
2. Preprocessing data,
3. Application of methods data mining,
4. Interpretation, evaluation and implementation the results to the pedagogical praxis.

Moodle logs activities including views and posts for all learning objects hosted in the system and provides „Reports" and statistics to help the content experts to improve the quality of eLearning courseware [27]. The records of the students' proceedings, created based on their activity in the course, however, do not contain information on the way of student's utilization of the material. The systems use log files to archive only data about the behavior of particular students in the course, which sources and activities s/he worked with, in what time periods, where from, etc. We can only find out whether the student has opened the material.

However, to get an idea of real transition of the students throughout the e-learning course, we need to consider several other important factors, one of them being the usage of implemented multimedia elements (e.g. interactive animations). All available electronic systems are able to record the time at which the student opened the website where the animation is situated and when he/she moved to another website. None of them, though, was concerned with the activity of the student from the point of view of manipulation with interactive media elements. Thus, the systems only stated the time that the students spent at the website where the media element was placed but the question if the student really worked with the element still remained unanswered. Therefore, it is only adequate to ask how to verify the activity/non-activity not only based on the transition throughout the course (opening the lesson, filling-in the quiz), but also via the detection of mouse movement or stating the interactivity of the student with the study material.

In the literature it is possible to meet with various attitudes to the definition of the concept multimedia. According to Neo and Neo [29]: *„medial elements can be differentiated as to the ability of perception (sentience) and control into text, graphic, animation, video and sound*". Rahman [30] extended this definition as follows: *„multimedia represent technology allowing for introducing text, sound, pictures, animations or video using interactive method*". In connection with learning styles Sonwalkar, however, uses 6 medial elements (the sixth being simulation), while these elements are interconnected by interactive aspects of learning [37]. Learning styles can be characterized according to Sonwalkar based on: L1 = apprenticeship; L2 = incidental; L3 = inductive; L4 = deductive; L5 = discovery.

If we reflect on interactivity as a medial element, it is interesting to read the statement of Shterev [34]: *„Any media element is presented by its start and also by its duration. It may be nominal, maximal and minimal. The starting time and duration define 2D temporal space*".

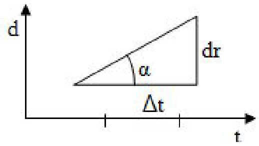


Fig. 3 Temporal-spatial dimension of medial element from the point of view of 2D temporal space [34]

Δt – time of duration of medial object
dr – real time of duration of the object
The correlation dr/Δt=tgα expresses the rate of speed of playback of the medial object, i.e. speed of reproduction.
Besides temporal-spatial dimension, interactivity can be evaluated also based on these three viewpoints:

1. Frequency (how frequently can the user react),
2. Possibilities of choice (how many choices can the user have at his disposal),
3. Significance (to what degree the decision of the user will influence the fact which will happen).

IV. Applying data mining techniques on the basis using interactive animations in LMS Moodle

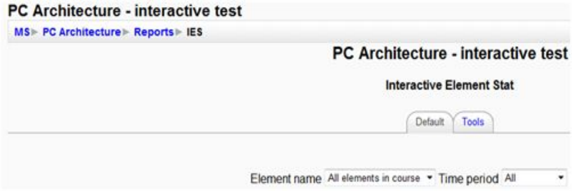
The reasons described above led us to develop a module which would, together with the original module "Reports", supply a complete report on student's activity even in cases when interactive media elements are implemented into the study material. We named it Interactive Element Stat (IES). The module was being developed since 2010 and was designed and programmed at the Department of Informatics at the Faculty of Natural Sciences, Constantine the Philosopher University in Nitra as the supporting system for the area of the analysis of educational activities of the students in LMS Moodle. Researching the current state of the issue, we found that no such kind of module has yet been developed, one that would be strictly aimed on evaluation of student's work with implemented interactive media element.
Main requirements on the module:

- the option of results display selection (whether the statistics is to be done for all the interactive media elements in the e-learning course or only for a particular interactive element),
- in statistics for the entire course, it is necessary to display a list of all the interactive course elements that were worked with, number of accesses to each of the elements, number of clicks, total time of student's work with the element,
- in statistics for a particular interactive element, it is necessary to display the name of the student who worked with the element, time spend on the work, student's way of manipulation with the element, what buttons were pressed, etc.,
- in statistics display, the option of time period selection

is necessary,

- the statistics has to enable export into MS Excel format for further processing of the data,
- the teacher has to be able to delete the created statistics and start gathering new data from the beginning.

The standard module in development Course Report was enhanced by a mod.php file, which includes an assigned reference to the module itself, which is displayed in the Module report list. The file is designated to control whether the user has sufficient privileges to display the reference, if so, the reference will be displayed. After clicking the link, the module itself will open, specifically the index.php file. This file represents the main screen of the graphic module with two tabs. The Default tab contains a form for choosing the type of statistics and the time period. The Tools tab contains two buttons to delete the recorded statistics (Figure 4).



PC Architecture - interactive test

MS ▸ PC Architecture ▸ Reports ▸ IES

PC Architecture - interactive test

Interactive Element Stat

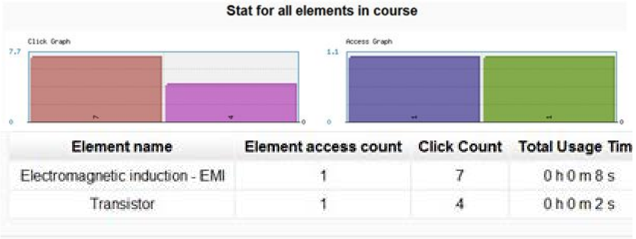
Element name: All elements in course
Time period: All

Fig. 4 Module Interactive Element Stat

The module form displays the two standard Select boxes. The first one contains a list of all the course elements that were a part of any activity. The second one is used to choose the time period of statistics record. The form can be submitted using the Submit button. An important part of the index.php file is a safety control, in which we determine whether, the user:

- is working with an existing course,
- is currently logged in,
- has sufficient privileges to work with the module.

The statistics display itself uses a table format; it is possible to display the main statistics for all the course elements using graphs. Designing and developing the module, we decided for two graph display, in which the first graph shows the number of interactions with a particular media element and the second one shows the number of accesses to the element in the framework of the whole course (Figure 5).



Stat for all elements in course

Element name	Element access count	Click Count	Total Usage Tim
Electromagnetic induction - EMI	1	7	0 h 0 m 8 s
Transistor	1	4	0 h 0 m 2 s

Fig. 5 Stat Results statistics and graphical evaluation Axis x: interactive element name, axis y: total number of accesses or the number of performed actions during the accesses.

The structure of the file exported to MS Excel is unified for all types of statistics provided by the module; it contains the *name of the statistics*, *statistic data from the table* and *informative foot* with the name of the module and time of the export itself.

V. Case study of application of the methods of data mining in the sphere of personalization of study material for the student

According to Romero, Ventura and García [33] Moodle does not provide a basic statistics module in which the teacher can obtain specific reports about detailed statistics about every single student's performance (how many hours on the site, how much time at every activity, etc.).

This problem we partially removed using module IES. Using information obtained from module IES we can detect more easily students with some learning problems, for example, students with a very low number of accesses and offer them a suitable learning style.

Therefore, we introduce so called association rules. Association rule mining is one of the most well studied mining methods [7]. Agrawal et al. [1] defined rules for techniques of data mining in this way: given a set of transactions, where each transaction is a set of items, an association rule is a rule of the form $X \rightarrow Y$, where X and Y are non-intersecting sets of items. Each rule is accompanied by two meaningful measures, confidence and support. Confidence measures the percentage of transactions containing X that also contain Y. Similarly, support measures the percentage of transactions that contain X or Y.

These rules of data mining have been applied to different learning management systems for building a recommender agent that could recommend on-line learning activities or shortcuts [17], for automatically guiding the learner's activities and intelligently generating and recommending learning materials [22], for determining which learning materials are the most suitable for students [21].

Rules are often applied to the whole system, which becomes adaptive at the personalization of the student, or to the particular part of the e-learning system, most frequently to the teaching part offering study material based on a suitably designed learning strategy (of the suitable learning style). In this case, the model FSLSM applied to the conditions of creating and providing the study material in the system Moodle is most frequently used.

The Learning Management System (LMS) Moodle enables teachers to create a lesson in form of a series of HTML pages. Lessons are created through the Lesson module or module Book. These modules are modules of third pages.

The module Book allows for simply creating multi-page texts, similarly as we are used to do it from printed books. We are not pressed to create many sources in HTML format of page; we can put all into one, thus at the same time increasing lucidity of the course. Possibility to create hierarchical structure of chapters and sub-chapters is also an asset [39].

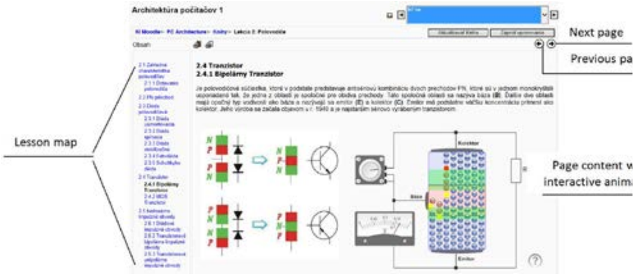


Fig. 6 Stat Interface and typical content page created by the module Book with implemented
interactive animation (Transistor), from original course in Slovak language

The teacher decides how many buttons will be on each content page and for each button what is the target page ("jump to"). The "Next page" button allows direct guidance of a student, i.e. he/she will follow the default path determined by the teacher. The other buttons along with map of the lesson allow the students to create their own path through the lesson [28]. The module Book is therefore not adaptive. For providing advanced adaptive behavior, we modified the original module and this module has name AdaptiveBook. This module enables the creation of lessons adapted to the learning styles of students according to FSLSM.

The module AdaptiveBook was used in e-learning course named Architecture of computers I, which is focusing on logic systems (winter semester of academic year 2013/2014, students of 1st year of the field Applied informatics). Students studying in the course were graduates of various secondary schools: secondary school of electrical engineering (15), grammar school (2), business academy (1), hotel academy (2).

In order to be able to match the suitable learning style to each student, ILS questionnaire was implemented into the module AdaptiveBook. At implementing the questionnaire we draw from experiences of [28], who implemented a module of similar character into LMS Moodle.

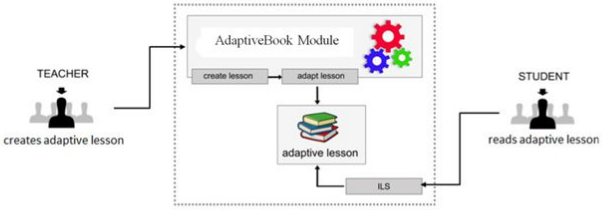


Fig. 7 Teacher's and student's interaction with AdaptiveBook module (Nakić, Graf and Granić, 2013)

By filling the ILS questionnaire out at the beginning of semester and applying association rules the student obtains personalization in the form of adaptive provision of the study material. By applying the module AdaptiveBook and the following continuous evaluation of study results during the semester we found out that in spite of the questionnaire filled-out and applying the association rules some students reached unsatisfactory study results. All these were the ones who did not graduate from secondary school of electrical engineering.

By monitoring the students´ activities using the standard configuration *Report* in Moodle it was not possible to determine their complete activity. Defining activity is an important step in personalization. Based on the information on the movement of the student in the course it is possible to apply association rules more consistently. On a regular basis, the module AdaptiveBook works on the basis of an allocated learning style to the particular student. But what if this style changes during the study due to unpredictable circumstances?

Based on access to the study materials we found out that students despite the provided learning style had problems with correct analyzing and understanding the provided study material. That is why they utilized very frequently a back transition to the previous parts of lessons, which was rather chaotic despite the module AdaptiveBook (Table 1).

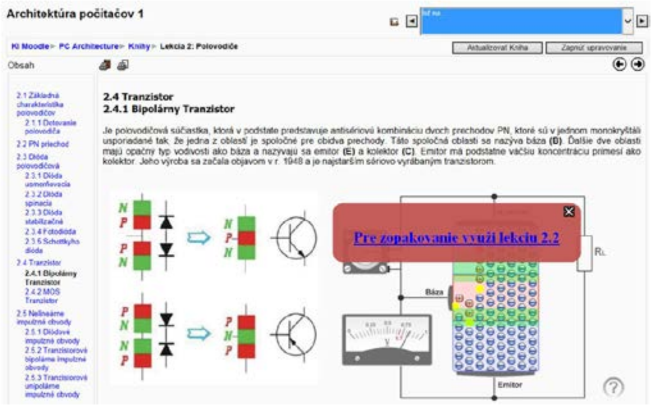
Table 1. Interactive matrix of transitions between individual lessons

	Start study	L1	L2	L3	L4	L5	L6	L7	L8	L9	End study
Start study	0	2450	852	356	124	258	689	346	734	428	45
L1	892	0	1987	556	87	2190	324	222	318	110	23
L2	634	1554	0	2041	918	796	369	567	216	257	51
L3	176	652	347	0	1321	821	221	705	599	375	74
L4	841	869	490	1458	0	1878	478	756	311	338	36
L5	654	512	1591	428	998	0	2887	568	850	151	111
L6	317	974	627	898	1370	350	0	1655	152	185	34
L7	268	498	623	495	580	915	1331	0	1201	100	174
L8	954	825	829	461	613	558	471	434	0	1637	190
L9	438	604	268	947	864	466	420	623	350	0	255
End study	526	249	315	185	277	216	265	170	57	46	0

The value in the column (above 1000) expresses the fact that the students realized the given activity most frequently and after it they continued with another activity with the highest maximal value situated in the nearest column. In case that there is more than one maximal value in the column of interaction matrix, it means that the student returned to this activity during the course of his study.

On this account we interconnected the IES module with the one of AdaptiveBook in order to identify the students´ activity more easily. By interconnecting the modules we obtained a tool, which allowed us to mine the data directly at activities of the students with the study material and propose continuous changes in the learning styles to them. These changes resulted in providing study materials, or offering a choice of its parts through implemented interactive animations. E-learning course has been considerably simplified. The amount of text and pictures decreased. They were replaced just by this type of interactive media element.

Connection of IES modules and AdaptiveBook allowed the students for fully utilizing interactive possibilities of implemented animations and finding one´s way in the study material by means of hyperlinks, which continually appeared in them. Provision of hyperlinks was realized based on the results of data mining from IES module and applying association rules. The prerequisite for the provision of hyperlinks was for example repeated utilization of one and the same function – a view of some of the animation parts, or a particular active work with animation, or its part.



Architektúra počítačov 1

KI Moodle ► PC Architecture ► Knihy ► Lekcia 2: Polovodiče

2.4 Tranzistor

2.4.1 Bipolárny Tranzistor

Je polovodičová súčiastka, ktorá v podstate predstavuje antisériovú kombináciu dvoch prechodov PN, ktoré sú v jednom monokryštáli usporiadané tak, že jedna z oblastí je spoločné pre obidva prechody. Táto spoločná oblasti sa nazýva báza **(B)**. Ďalšie dve oblasti majú opačný typ vodivosti ako báza a nazývajú sa emitor **(E)** a kolektor **(C)**. Emitor má podstatne väčšiu koncentráciu primesí ako kolektor. Jeho výroba sa začala objavom v r. 1948 a je najstarším sériovo vyrábaným tranzistorom.

Pre zopakovanie využi lekciu 2.2

Fig. 8 Adaptive provision of hyperlinks by the module AdaptiveBook and module IES, from original course in Slovak language

After a repeated analysis of accesses at the end of the semester and setting up of interaction matrix of the transition of the students through e-learning course we found out that students passed the course fluently. It appears from this that a suitable learning style for each individual was chosen and the study material was adjusted to the possibilities and abilities of every student.

Table 2. Interaction matrix of transitions between individual lessons (after the modification by IES module and AdaptiveBook)

	Start study	L1	L2	L3	L4	L5	L6	L7	L8	L9	End study
Start study	0	1609	720	662	792	558	822	521	880	830	43
L1	955	0	1595	435	849	874	729	912	420	218	62
L2	949	674	0	1455	896	411	862	325	930	538	30
L3	958	449	174	0	1355	297	808	375	358	994	42
L4	221	706	667	721	0	1279	831	480	814	266	53
L5	551	656	742	505	378	0	1004	458	609	292	20
L6	156	795	302	804	928	429	0	1108	351	203	175
L7	663	694	251	846	956	892	676	0	1184	161	42
L8	982	356	826	703	629	710	615	123	0	1523	262
L9	519	546	334	590	495	554	863	294	1187	0	358
End study	259	177	142	236	112	190	127	90	134	217	0

VI. Conclusion

According to Felder and Silverman [12], active learners are comfortable with problem-solving activities and group discussions, they prefer answering questions and doing exercises but less theory and examples. In contrary, reflective learners learn by reflecting on the matter and thinking things through.

To determine the learning strategy (learning style) is not a simple process. In the contribution we gave an example that despite filling out the structured questionnaire ILS the allotted learning style to the student need not necessarily suit him during the whole semester. At present, authors of professional publications dealing with the implementation of ICT in education [3] point to the fact that the development of ICT is higher than their actual use, and requires thinking about the elements that we need to improve to produce ICT effective integration in educational processes [24]. As Internet use has proliferated, e-learning systems have become increasingly popular. Many researchers have taken a great deal of effort to promote high quality e-learning environments, such as adaptive learning environments, personalized/adaptive guidance mechanisms, and so on. These researches need to collect large amounts of behavioral patterns for the verification and/or experimentation. However, collecting sufficient and correctly behavioral patterns usually takes a great deal of time and effort [9].

References

[1] R. Agrawal, T. Imielinski, A. Swami, “Mining association rules between sets of items in large databases,” in *Proceedings of the ACM SIGMOD international conference on management of data*, Washington DC., USA, pp. 207–216, (1993).

[2] Z. Balogh, M. Magdin, T. Turčáni, M. Burianová, “Interactivity elements implementation analysis in e-courses of professional informatics subjects, ” in *Efficiency and Responsibility in Education 2011* : 8th International Conference, pp. 5-14, (2011).

[3] W. Bhuasiri, O. Xaymoungkhoun, H. Zo, J.J. Rho, A.P. Ciganek, “Critical success factors for e-learning in developing countries: A comparative analysis between ICT experts and faculty, ” in *Computer & Education*, vol. 58, no. 2, pp. 843-855, (2012).

[4] P. Brusilovsky, “From Adaptive Hypermedia to the Adaptive Web,” in *Mensch & Computer*. Interaktion in Bewegung. Stuttgart: B. G. Teubner, pp. 21-24, (2003).

[5] J. Burgerová, “Internet vo výučbe a štýly učenia,” Prešov: SAMO AUTOMATION, (2001).

[6] M. Cater, “Incorporating Learning Styles into Program Design,” in [online]. LSU AgCenter ODE Blog, (2011).

[7] A. Ceglar, J. Roddick, “Association Mining,” in *ACM Computing Surveys*, 38(2), pp. 1-42, (2006).

[8] Y.C. Chang, W.Y. Kao, C.P. Chu, “A Learning Style Classification Mechanism for E-learning,” in *Computers and Education*, vol. 53, pp. 273-285, (2009).

[9] Y.C. Chang, Y.C. Huang, C.P. Chu, “B2 model: A browsing behavior model based on High-Level Petri Nets to generate behavioral patterns for e-learning,” in *Expert Systems with Applications*, vol. 36, no. 10, pp. 12423-12440, (2009).

[10] F. Coffield, D. Moselea, E. Hall, K. Ecclestone, “Learning styles and pedagogy in post – 16 learning, “ in *A systematic and critical review*. London: Learning and Skills Research Centre, pp. 182, (2004).

[11] A.D. Cohen, A S. J. Weaver, “Styles and strategies-basedinstruction: A teachers’ guide,“ in Minneapolis, MN: Center for Advanced Research on Language Acquisition, University of Minnesota, pp. 200, (2006).

[12] R.M. Felder, L.K. Silverman, “Learning and Teaching Styles in Engineering Education,“ in *Engineering Education* 78(7), pp. 674–681, (1988).

[13] R.M. Felder, B.A. Soloman, “Index of learning styles questionnaire,“ (1997), http://www.engr.ncsu.edu/learningstyles/ilsweb.html, (retrieved April 01, 2014).

[14] P. Honey, A. Mumford, “The Manual of Learning Styles,“ in *Peter Honey Publications*, Maidenhead, (1992).

[15] M. Houška, M. Houšková Beránková,“ Pedagogical Efficiency of Multimedia Lectures on Mathematical Methods in Economics,” in *Proceedings of the 7th International Conference on Efficiency and Responsibility in Education* (ERIE 2010), Prague, pp. 94-101, (2010).

[16] H.Y. Jeong, C.R. CHoi, Z.J. Song, “Personalized Learning Course Planner with E-learning DSS Using User Profile,” in *Expert Systems with Applications*, vol. 39, pp. 2567-2577, (2012).

[17] J. Kapusta, M. Munk, M. Turčáni, “Experimental comparison of adaptive links annotation technique with adaptive direct guidance technique,“ in *Webist 2009 : ACM Conference Proceedings*. 5th International Conference on Web Information Systems and Technologies, Lisboa Portugal, 23.-26. March 2009. - Lisabon : Insticc Press, pp. 250-256, (2009).

[18] M.J. Koehler, P. Mishra, “Introducing TPCK,” in J. A. Colbert, K. E. Boyd, K. A. Clark, S. Guan, J. B. Harris, M. A. Kelly & A. D.

Thompson (Eds.), Handbook of Technological Pedagogical Content Knowledge for Educators, New York: Routledge, pp. 1–29, (2008).

[19] D.A. Kolb, “Experiential learning: Experience as the source of learning and development,” Englewood Cliffs, NJ: Prentice Hall, pp. 288, (1984).

[20] K. Kostolányová, J. Šarmanová, O. Takács, “Structure of study supports for adaptable instruction,” in *The New Educational Review*, 25(3), pp. 235-247, (2011).

[21] K. Kostolányová, O. Takács, J. Šarmanová, “Adaptive Education Process Modeling,” in *Proceedings of the 10th International Conference on Efficiency and Responsibility in Education 2013*, pp. 300-308, (2013).

[22] J. Lu, “Personalized e-learning material recommender system,” in *International conference on information technology for application*, Utah, USA, pp. 374–379, (2004).

[23] J. Mareš, “Styly učení žáků a student,” Praha: Portál, pp. 239, (1988).

[24] J.M.J. Melia, J. Gonzales-Such, M.R. Garcia-Bellido, “Evaluative Research and Information and Communication Technology (ICT),” in *Revista Espanola de Pedagogia*, 70(251), pp. 93-110, (2012).

[25] P. Mishra, M.J. Koehler, “Technological pedagogical content knowledge: A framework for teacher knowledge,“ in *Teachers College Record*, 108(6), pp. 1017-1054, (2006).

[26] J. Mostow, J. Beck, H. Cen, A. Cuneo, E. Gouvea, C. Heiner, C., “An educational data mining tool to browse tutor-student interactions: Time will tell!,” in *Proceedings of the Workshop on Educational Data Mining*, Pittsburgh, USA, pp. 15–22, (2005).

[27] K. Nagi, P. Suesawaluk, “Research analysis of Moodle reports to gauge the level of interactivity in elearning courses at Assumption University, Thailand,” in *Proceedings of the International Conference on Computer and Communication Engineering*, Kuala Lumpur, Malaysia, May 13-15, pp 772-776, (2008).

[28] J. Nakić, S. Graf, A. Granić, “Exploring the adaptation to learning styles: The case of AdaptiveLesson module for Moodle,” in *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7946 LNCS, pp. 534-550, (2013).

[29] T.K. Neo, M. Neo, “Integrating multimedia into the Malaysian classroom: Engaging students in interactive learning,” in *The Turkish Online Journal of Educational Technology*, 3(3), pp. 31-37, (2004).

[30] S.M. Rahman, K.N. Tsoi, G. Dettrick, “Multimedia as an educational tool: An overview and the future,” in *Proceedings of the Third International Interactive Multimedia Symposium*. Monash University, pp. 328-33, (1996).

[31] P. Robinson, “Individual differences and instructed language learning,” Amsterdam: John Benjamins, pp. 387, (2002).

[32] C. Romero, S. Ventura, “Data mining in e-learning,” Southampton, UK: Wit Press, (2006).

[33] C. Romero, S. Ventura, E. García, “Data mining in course management systems: Moodle case study and tutorial,” in *Computers and Education*, 51 (1), pp. 368-384, (2008).

[34] J. Shterev, “Modeling of Interaction on Multimedia Streams and Objects by Application of Petri Nets,” in *International Conference on Computer Systems and Technologies - CompSysTech' 2005*. Varna, Bulgaria : University of Rouse, pp. IIIB.21-1 - IIIB.21-6, (2005).

[35] L.S. Shulman, “Those who understand: Knowledge growth in teaching,” in *Educational Researcher*, 15(4), (1986).

[36] I. Šimonová, P. Poulová, “Students' Feedback after Studying Online Courses Reflecting Individual Learning Styles,“ in *Proceedings of the 2013 International Conference on Information, Business and Education Technology (ICIBET)*. Book Series: *Advances in Intelligent Systems Research*, Volume 26, pp.: 971-975, (2013).

[37] N. Sonwalkar, “A New Methodology for Evaluation: The Pedagogical Rating of Online Courses,” in *Campus Technology from Syllabus Media Group*, 15(6), pp. 18-21, (2001).

[38] N. Stefanovic, D. Stefanovic, B. Arsovic, “Adaptivity in e-learning LMS platform,” in *Metalurgia International*, 18 (3), pp. 156-162, (2013).

[39] P. Švec, “Možnosti rozšírenia LMS Moodle o ďalšie moduly,“ in *Technológia vzdelávania: Slovenský učiteľ*, 15(5), pp. 6-8, (2007).

[40] O. Takács, J. Šarmanová, K. Kostolányová, “Analysis of learning styles for adaptive E-learning,” in *Communications in Computer and Information Science*, 188 CCIS (PART 1), pp. 368-376, (2011).

[41] I. Turek, “Učebné štýly a rozvoj schopností žiakov učiť sa,” Banská Bystrica : Metodicko-pedagogické centrum, pp. 40, (2002).

An Approach Based on Reinforcement Learning for Quality of Experience (QoE) Control

F. Cimorelli, M. Panfili, S. Battilotti, F. Delli Priscoli, C. Gori Giorgi, and S. Monaco

Abstract**—** The paper proposes an approach based on Reinforcement Learning algorithms to cope with Quality of Experience (QoE) Control in Future Internet networks. The proposed approach aims at guaranteeing the satisfaction of personalized Quality of Experience (QoE) requirements to the applications. The paper refers to the Future Internet architecture which has been developing in the framework of the PLATINO and FI-WARE projects.

Keywords**—** Quality of Experience; Quality of Service; Future Internet; Reinforcement Learning.

I. Introduction

THE FI-WARE UE FP7 project [1] and the PLATINO Italian National project [2] are two major projects which are trying to address the issues raised by the design of the Future Internet and, in particular, by the necessity to assure a personalized Quality of Experience (QoE) which represents a key Future Internet novelty.

In the authors' vision (see also [28]), the Future Internet overall target is to allow Applications to transparently, efficiently and flexibly exploit the available network resources, aiming at achieving a satisfaction level meeting the personalized users' QoE [3], [4]. The International Telecommunication Union (ITU-T) defines QoE as: *The overall acceptability of an application or service, as perceived subjectively by the end-user* [5]. In other words, QoE is the perception that the user has about the performance of the network when he uses an application and about how this application is usable.

A large amount of research is on-going in the field of the identification of the personalized user expected QoE level in a given context for a given application (e.g. see [6], [7] for voice and [8], [9] for video applications, respectively), as well as of the functions for QoE computation, including passive and active monitorable feedback parameters which serve as independent variables for these functions; in particular, several works focus on studying the QoE relation with network QoS parameters [10]. By *passive* and *active parameters* we mean the ones which are independent of and dependent on the active involvement of the users in their computation, respectively [11].

This paper focuses (i) on the Future Internet *cognitive* architecture supporting QoE, (ii) on a flexible way of computing QoE on the basis of a suitable set of (passive and/or active) monitorable parameters, and (iii) on the definition of QoE Agents which, for each application, perform control functions aiming at minimizing the difference between the computed QoE and the desired QoE level.

A first key innovation of the paper is that, for a given application, the structure of the proposed QoE Agent and its way of working are flexible, since they do not impose specific requirements to the function for QoE computation and to the set of monitorable (passive and/or active) parameters.

A further fundamental innovation of the paper is that the QoE Agents, on the basis of the monitorable parameters, aim at approaching the desired QoE level of the applications by dynamically selecting the most appropriate class of service supported by the network. In this work, the selection is driven by an adaptive algorithm based on the Reinforcement Learning (RL) methodology. The proposed dynamic approach differs from traffic classification approaches found in the literature (e.g., [12] and references in [13]), based on host-level communication behaviour-based approaches, or on statistical approaches relying on data mining methodologies, since they statically determine the class of service of the application.

II. Future Internet Core Platform Architecture

This section gives a high level overview of the Future Internet Core Platform architecture, built on the work in [4], [14], [15], [16], [28]. Figure 1 highlights some key functionalities of the Future Internet Core Platform [17]. Such functionalities can be implemented by means of distributed Agents to be transparently embedded in properly selected network nodes (e.g., Mobile Terminals, Base Stations, Backhaul Network entities, Core Network entities).

The *Sensing and Data Processing functionalities* are in charge of the *monitoring* and the preliminary *filtering* of properly selected possibly heterogeneous information (e.g., including device, network performances, user profiles, network provider policies, etc.).

This work was supported in part by the EU project FI-WARE (Grant Agreement n° 285248), and partly from the Italian project PLATINO (Grant Agreement n° PON01_01007).

F. Cimorelli, M. Panfili, S. Battilotti, F. Delli Priscoli, C. Gori Giorgi and S. Monaco are with the Department of Computer, Control and Management Engineering “Antonio Ruberti” of the University of Rome “La Sapienza”, Rome, Italy (e-mail: {dellipriscoli, panfili}@dis.uniroma1.it).

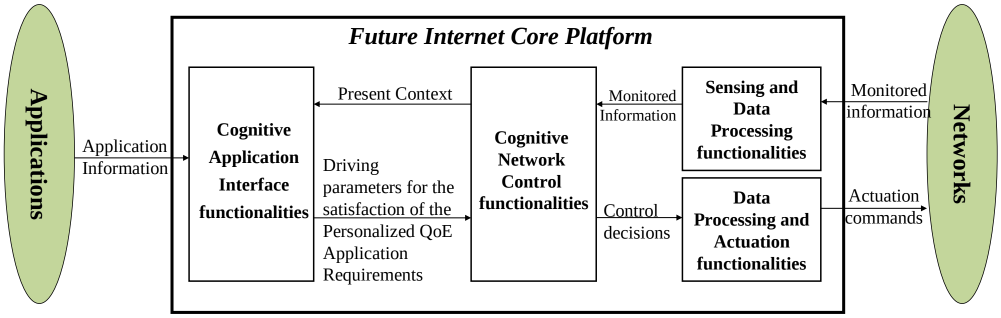

Future Internet Core Platform
Applications
Application Information
Cognitive Application Interface functionalities
Present Context
Driving parameters for the satisfaction of the Personalized QoE Application Requirements
Cognitive Network Control functionalities
Monitored Information
Control decisions
Sensing and Data Processing functionalities
Data Processing and Actuation functionalities
Monitored information
Actuation commands
Networks

Figure 1: Future Internet Core Platform concept

The *Cognitive Network Control functionalities* (in the following, also simply referred to as *Network Control* functionalities") consist of a set of cooperative, technology-independent algorithms and procedures which are in charge of the formal description of the Monitored Information in homogeneous metadata, as well as of the proper aggregation of these metadata to form a multi-layer, multi-network *Present Context.* In addition, the Network Control functionalities, on the basis of the Present Context and on the above-mentioned driving parameters are in charge of taking control decisions concerning specific Network Control problems.

The *Cognitive Application Interface* works on the basis of (i) the Present Context, (ii) parameters gathered from the application which can include direct or indirect user feedbacks. The Cognitive Application Interface, among the others, include two key agents, namely the *QoE Evaluator* and the *QoE Controller.*

The QoE Evaluator, is in charge of evaluating, for each in progress application, the personalized QoE expected by the user (hereinafter referred to as *Target QoE*) and the actual, present QoE experienced by the user (hereinafter referred to as *Perceived QoE*).

The QoE Controller is in charge of computing, for each in progress application, the Driving Parameters that should drive the Network Control functionalities to take the control decisions aiming at reducing the difference between the Target QoE and the Perceived QoE (namely the so-called *QoE Error*), as well as at optimizing the exploitation of the network resources ([27]). The Driving Parameters relevant to a given application can include, among others, QoS aspects (e.g., maximum tolerated delay, minimum throughput to be guaranteed, etc.), security aspects (e.g., desired encryption, allowed network nodes, etc.) or content/service aspects (e.g. the most appropriate service/content mix, etc.). In general, the Driving Parameters are multi-layer and multi-network.

Finally, the *Data Processing and Actuation functionalities* are in charge of "translating" the technology-independent control decisions taken by the Network Control functionalities, in technology-dependent actuation commands which put into operation on the Networks the above-mentioned decisions.

This paper focuses on the Cognitive Application Interface and, mainly, on the QoE Controller. Instead, the Cognitive Network Control functionalities and the QoE Evaluation are outside the scope of this paper; instances of such functionalities can be found in admission control [17], [25], [26], routing [19], [34], [36], congestion control and scheduling [20], dynamic capacity assignment [21], [22], medium access control [23], load balancing [35].

III. THE QOE CONTROLLER

The QoE Controller has to deduce the Driving Parameters aiming at the satisfaction of the Personalized QoE Application Requirements, namely at the minimization, for each in progress application of its QoE Error, defined as the difference between the Target QoE and the Perceived QoE of the application in question. To reach this goal, the QoE Controller should know – or, at least, estimate – the correlation between its decisions (the selected driving parameter) and the Perceived QoE in a given Present Context. In this respect, no model of the traffic flows in the network can be assumed, since the network behaviour depends on too many factors: traffic characteristics of the on-going applications, network topologies, network resource management algorithms, congestion control algorithms, and so on. The decision strategy must therefore be learned on-line by trial and errors. In this respect, this paper proposes Reinforcement Learning as the key technology to enable an organized on-line exploration of the possible decision strategies, named policies, and the exploitation of the best policy to be enforced.

The QoE Controller can be implemented by means of Agents (referred as QoE Agents) to be carefully embedded in properly selected network nodes (e.g., Base Stations and Mobile Terminals in a wireless environment).

In this paper we present two algorithms. The first one, referred to as single-agent learning, proposes that the decisions (i.e., the value of the Driving Parameters) are taken by each Agent on the basis of its local knowledge of the Present Context and of the so-called Status Signal, which represents in a concise way the overall Network status, broadcast by a single centralized entity, named Supervisor Agent. In the second algorithm, referred to as multi-agent learning, the Agents communicate their QoE Error to the Supervisor Agent, which computes and broadcasts the decisions; the relevant problem can be modeled as a Multiagent System [32].

In both algorithms the learning approach consists in a model-free adaptive feedback approach: the effect of the decisions are observed as a variation of the QoE Error, and the decisions are taken based on past-observations. Reinforcement Learning (RL) is an interesting approach to solve both single and multiple agents problem: RL focuses the attention on learning by the individual from directly interaction with its environment, without relying on complete model of the environment. The interaction, between agent and environment, is defined by formal framework (states, actions, rewards or costs) and the environment is typically formulated as a finite-state Markov Decision Process.

Both approaches entail the presence of a centralized entity, which sends control signalling to the Agents. This approach is well-matched to the current trends in managing communication network, as with the Software Defined Network [24].

Concerning the Driving Parameters, in case the underlying networks can be driven by a set of QoS parameters, the QoE Controller has to decide the most appropriate target value of each parameter to drive the Perceived QoE as close as possible to the Target QoE. Since the control action has a large number of degree of freedom, the solution space exploration may take a large amount of time: so, the QoE Controller task may be complex. A simpler control task arises if QoS management of the underlying network is organized in Classes of Service (CoS). In this case, the role of the QoE Controller is to select the most appropriate CoS for the on-going applications (i.e. the Driving Parameters directly determine the CoS of the applications) aiming at reducing the QoE Error.

A. Single Agent Reinforcemente Learning

The problem is described by a Markov Decision Process, a tuple $\{X, A, pr, r\}$, where X is the finite state space, A is the finite set of agent actions, pr is the transition probability function, r is the one-step reward function. The state $x \in X$, that describes the environment, can be altered by the agent action $a \in A$. The environment changes state according to the state transition probabilities given by $pr(x, a, x')$. The reward evaluates the immediate effect of action a. The behavior of the agent is described by its policy π, which specifies how the agent chooses its actions given the state, it may be either stochastic, $\pi: X \times A \to [0,1]$, or deterministic, $\pi: X \to A$.

We consider a common reinforcement learning technique, known as Q-Learning [29], [30], that works by learning the action-value function. The action-value function $Q^{\pi}(x, a)$ is the expected return starting from x, taking action a, and thereafter following policy π; it satisfies the Bellman equation:

$$Q^{\pi}(x,a) = \sum_{x' \in X} pr(x,a,x')[r(x,a,x') + \gamma \max_{a' \in A} Q^{\pi}(x',a')] \quad (1)$$

where the discount factor $\gamma \in [0,1)$ weights immediate rewards versus delayed rewards.

Let $Q^{*}(x, a)$ be the optimal action-value function, defined as:

$$Q^{*}(x,a) = \max_{\pi} Q^{\pi}(x',a'), \forall x \in X, a \in A(x) \quad (2)$$

Then, the agent, computing $Q^{*}(x, a)$, can maximize its long-term performance, while only receiving feedback about its immediate, one-step performance. The greedy policy is deterministic and picks for every state the action with the highest Q-value:

$$\pi(x) = \arg\max_{a' \in \mathrm{A(x)}} Q(x,a') \quad (3)$$

The Q-learning approach derives the policy on-line by estimating the (action, state)-values with the following update rules:

$$Q(x,a) \leftarrow$$
$$(1 - \alpha(x))Q(x,a) + \alpha(x)[r(x,a,x') + \gamma \max_{a' \in \mathrm{A}} Q(x',a')] \quad (4)$$

where the learning rate $\alpha(x) \in [0,1]$ determine the convergence speed and accuracy.

B. Multiagent Reinforcement Learning

The generalization of the Markov Decision Process to the multiagent case is a stochastic game (SG) described by a tuple $\{X, A_1, \ldots, A_N, pr, r_1, \ldots, r_N\}$ where N is the number of agents, X is the discrete set of environment states, A_n is the discrete sets of actions available to the agent $n, n = 1, \ldots, N,$ yielding the joint action set $A = A_1 \times \ldots \times A_N$, $pr: X \times A \times X \to [0,1]$ is the state transition probability function, and $r_n: X \times A \times X \to \mathbb{R}$ is the reward functions of the agent $n, n = 1, \ldots, N$. The state transitions and the reward depend on the joint action of all the agents, $\mathbf{a} = [a_1^{\mathrm{T}}, \ldots, a_N^{\mathrm{T}}], \mathbf{a} \in A, a_n \in A_n, n = 1, .., N$. The policies $\pi_n: X \times A_n \to [0,1]$ form together the joint policy $\boldsymbol{\pi}$. Clearly, the Q-function of each agent n (Q_n^{π}), depends on the joint action and is conditioned on the joint policy:

$$Q_n(x, a_1, \ldots, a_N) =$$
$$\sum_{x' \in X} pr(x, a_1, \ldots, a_N, x')[r_n(x, a_1, \ldots, a_N, x') +$$
$$\gamma Q_n(x', \pi_1, \ldots \pi_N)], n = 1, \ldots N \quad (5)$$

where $Q_n(x', \pi_1, \ldots \pi_N)$ is a weighted sum of $Q_n(x, a_1, \ldots, a_N)$.

Considering the single agents Q-learning approach (4), it is possible to define an analogue approach for Multiagetn RL as follow:

$$Q_n(x, \mathbf{a}) \leftarrow (1 - \alpha(x))Q_n(x, \mathbf{a}) + \alpha(x) \Big[r_n(x, \mathbf{a}, x') +$$
$$\gamma \,\mathrm{eval_n} \Big(\boldsymbol{\pi}(x') Q\big(x', \pi_n(x')\big)\Big)\Big], n = 1, \ldots N \quad (6)$$

$$\boldsymbol{\pi}(x) = \mathrm{solve}_{\pi}\big(Q_1(x, \mathbf{a}), \ldots, Q_N(x, \mathbf{a})\big) \quad (7)$$

where solve_{π} is a selection mechanism mapping from one stage games into joint distributions and $\mathrm{eval_n}$ gives the expected return of agent n given this joint distribution.

Littman in [31] presents a convergent algorithm, denoted friend-or-foe Q-learning (FFQ), that, in fully cooperative SG (e.g. $r_1 = \cdots = r_N$) or fully competitive SG (i.e. $r_1 = -r_2$), converges to the value Nash-Q [32]. Furthermore, in fully cooperative SG, if a centralized controller were available, the task would reduce to a Markov decision process (the action space would be the joint action space of the SG) and the goal could be achieved by learning the optimal joint-action values with simple Q-learning:

$$Q(x,\mathbf{a}) \leftarrow \\ \big(1-\alpha(x)\big)Q(x,\mathbf{a})+\alpha(x)[r(x,\mathbf{a},x')+\gamma \max_{\mathbf{a}'\in \mathrm{A}} Q(x',\mathbf{a}')] \qquad (8)$$

C. Problem statement

A generic network with the following features is considered:

1) available link capacity, denoted with B_{link};
2) M Application types, each one characterized by an average transmission bitrate $b_m,\ m = 1,\dots,M$;
3) N end-nodes/agents, each one supporting one particular application and characterized by personalized Target QoE level denoted $TQoE_n,\ n = 1,\dots,N$.

It is assumed that the network supports C classes of service. At each time step t, each agent n selects the most appropriate service class to be associated with the application supported by the node in question. We define $a_n(t\,),\ n = 1,\dots,N$, the control action of node n at time step t. Let $\mathbf{a}(t)$ be the vector of control action of all nodes, i.e.:

$$\mathbf{a}(t) = \big(a1(t),\dots,a_N(t)\big), \text{ where } a_n(t) \in \{1,\dots C\} \qquad (9)$$

The control objective is to minimize the error between the measured Perceived QoE, denoted $PQoE_n$, and the QoE target, for each node n.

IV. QoE Control Algorithms

Two multi agent RL approaches are proposed to solve the problem defined in the previous section. In both approaches, a soft method can be considered in order to address the exploration problem. In particular, ε-greedy policy is a soft method that consists in the selection of a random action with a small probability; in details, it selects: i) with probability $1-\varepsilon$, the greedy action (7), and ii) with probability ε, a random action $a \in A$, where the parameter $\varepsilon \in [0,1]$ weights the exploration of the state-space versus the exploitation of the current estimates of the (action,state)-values.

A. Single Agent Reinforcement Learning Approach

In the single-learning algorithm, at each decision period each Agent tries to minimize its QoE error by deciding its CoS for the next time interval, based on the local feedback on the available transmission rate, and on the Status Signal, which communicates the number of Agents which currently opted for each CoS. The decision is based on the estimate of the expected QoE error which may be achieved by switching to a given CoS. In this approach the single agent Q-learning is directly applied applied to the multi agent case, thus the joint actions are not consider. In order to model all information to solve the problem, we define the following Markov decision processes $\{X,A,pr,r_n\}$, for each agent n:

1) The space state X describe the environment; considering that, the state $x(t)$ represents the vector of active nodes enjoying the service m, $m = 1,\dots,M$, using the class of service $c, c = 1,\dots,C$, at time t: $x(t) = \big(n_{11}(t),\dots,n_{1C}(t),\dots,n_{M1}(t),\dots,n_{MC}(t)\big)$, where $n_{cm} = 0,1,\dots;\ c = 1,\dots,C;\ m = 1,\dots,M$. Thus the finite state space is defined as $X = \{x = (n_{mc}), c = 1,\dots,C; m = 1,\dots,M\}$
2) The action set represents, for each agent, the class selected for the transmission: $A_n = A = \{1,\dots,C\}, n = 1,\dots,N$
3) pr is the transition probability function;
4) For each agent n the cost $r_n(x,a,x')$ is defined by the error between the Perceived QoE $PQoE_n(x,a,x')$, and the Target QoE of agent n, $TQoE_n$: $r_n(x,a,x') = |\,PQoE_n(x,a,x') - TQoE_n|,\ n = 1,\dots,N$

In this case each agent solves an independent Q-learning algorithm, thus from (4):

$$Q_n(x,a) \leftarrow \big(1-\alpha(x)\big)Q_n(x,a)+\alpha(x)[r_n(x,a,x') + \\ \gamma \max_{a'\in \mathrm{A}} Q_n(x',a')] \qquad (10)$$

B. Multiagent Reinforcement Learning: Friend Q-Learning

In the multi-learning algorithm, the Supervisor Agent tries to minimize the average square QoE Error of the Agents by deciding their CoS for the next time interval. The decision is based on the estimate of the expected average square QoE error which is achieved by switching to a given CoS; the estimates are updated based on the QoE error measures sent by the Agents.

In this approach a static game is considered, it means a SG with $X = \emptyset$, in which the reward depends only on the joint actions. In particular we consider the following static game $\{A_1,\dots A_N, r_1,\dots,r_N\}$ where N is the number of agents, $A_1 = \dots = A_N = \{1,\dots C\}$ are the discrete sets of actions available to the agents, yielding the joint action set $A = A_1 \times \dots \times A_N = \{\mathbf{a} = [a_1^{\mathrm{T}},\dots,a_N^{\mathrm{T}}], \mathbf{a} \in A, a_n \in A_n, a = 1,\dots N\} = \{1,\dots,C\}^N$ and $r_n\colon A \to \mathbb{R}, n = 1,\dots N$ are the cost functions of the agents. For each agent n the cost $r_n(\mathbf{a})$ is defined by the error between the Perceived QoE $PQoE_n(\mathbf{a})$, that the agent n achieves when the joint action $\mathbf{a}$ is taken, and the Target QoE of agent n, $TQoE_n$:

$$r_n(\mathbf{a}) = |\,PQoE_n(\mathbf{a}) - TQoE_n|, n = 1,\dots,N \qquad (11)$$

Thus, the MARL approach could be described by the following equation derived by eq. (6) and eq. (7):

$$Q_n(\mathbf{a}) \leftarrow (1-\alpha)Q_n(\mathbf{a}) + \alpha\big[r_n(\mathbf{a}) + \gamma\, \mathrm{eval_n}\big(\boldsymbol{\pi} Q(\pi_n)\big)\big] \qquad (12)$$

$$\boldsymbol{\pi} = \mathrm{solve}_\pi\big(Q_1(\mathbf{a}),\dots,Q_N(\mathbf{a})\big) \qquad (13)$$

where solve_π returns a particular type of equilibrium and $\mathrm{eval_n}$ gives the expected return of agent n given this equilibrium.

Friend Q-learning approach converges if the SG has at least one coordination equilibrium. The coordination equilibrium is a particular Nash equilibrium, in which all players achieve their highest possible value:

$$r_n(\pi_1, \dots \pi_N) = \max_{a_1 \in A_1, \dots, a_N \in A_N} r_n(a_1, \dots, a_N), n = 1, \dots, N \quad (14)$$

If the SG is fully cooperative (e.g. $r_1 = \dots = r_N$) then there is at least one coordination equilibrium. Thus, in order to guarantee the convergence to coordination equilibrium, it is necessary to modify the SG definition provided in the previous section such that $r_1 = \dots = r_N$. One possible way to modify the static game $\{A_1, \dots A_N, r_1, \dots, r_N\}$ is to consider a new cost function:

$$r'_1 = \dots = r'_N = f(r_1, \dots, r_N); \quad (15)$$

where r_n is defined in eq. (11) and an example of function f could be the Euclidean norm ($f = \|\cdot\|_2$).

Considering the static fu
$\{A_1, \dots A_N, r'_1, \dots, r'_N\}$, the Friend Q-lea

$$Q_n(\mathbf{a}) \leftarrow (1 - \alpha)Q_n(\mathbf{a}) + \alpha[r'_n(\mathbf{a})$$

Note that $Q_1 = \dots = Q_N$, thus
entity, the original SG problem c
decision process:

$$Q(\mathbf{a}) \leftarrow (1 - \alpha)Q(\mathbf{a}) + \alpha[r'(\mathbf{a}) +$$

V. SIMULATION I

Numerical simulations were p
effecttiveness of the proposed app
were modeled to illustrate the poten
algorithms, it is assumed that the
CoSs, a small number of Agents, n
considered. The CoS is selected d
algorithms and the decisions oc
intervals. The network has a c
sources (the Agents) transmitting t
packets to a second switch, which
destinations. The link between the
characterized by the available link c

The network supports three diffe
and three different types of applic
characterized by an average transm

Each Agent has its specific
computed based on the well-known

$$f_{QoE} = \alpha \cdot e^{-\beta \cdot p_{QoS}} + \gamma$$

where α, β and γ are parameters to be tuned depending on the Application, in this case $\alpha = 1$, $\beta \in \{0.5, 0.7, 1\}$ depending on the application, and $\gamma = 0$. The IQX hypothesis is formulated with QoS as parameter (p_{QoS}), reflecting the objective service quality, in this case p_{QoS} is the error between requested and allocated bandwidth.

We consider different simulation environments, in each one each end-node is characterized by a personalized target QoE level E_n, $n = 1, \dots, 8$, a random value taken in the set $\{0.7, 0.8, 0.9\}$. During the simulation, each scenario is characterized by a given association between the Target QoE level and the type of application, characterized by the average transmission bitrate and denoted $b_n \in \{0.6, 1.2, 2\}, n = 1, \dots, 8$, furthermore the particular association between agent and type of application influences the scenario offered traffic load, computed as sum of average transmission bitrate of each end-node: $\eta_{off} = \sum_{n=1}^{N} b_n$.

To evaluate the overall algorithm performance in different traffic conditions, each scenario were simulated with three different value of available link capacity, in particular: $0.7\eta_{off}, 0.8\eta_{off}$ and $0.9\eta_{off}$, denoted, respectively, *High*, *Medium* and *Low* traffic condition.

The CoS of an application is dynamically chosen according

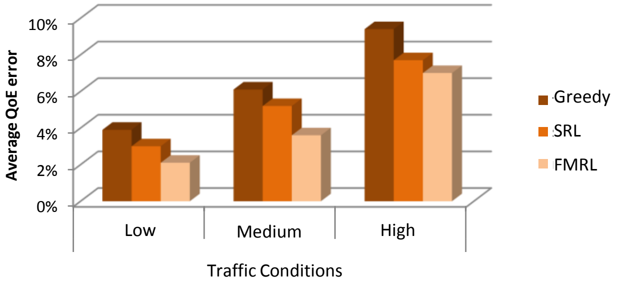


Figure 2 - Simulation results: QoE error per traffic load condition.

The obtained results show that the multi-learning algorithm outperforms the single-agent one, which, in turn, outperforms the fixed policy. However, it should be noted that the amount of exchanged information is larger in the multi-agent algorithm than in the single-agent one, whereas it is null, during the Application lifetime, in the static CoSs case. The evaluation of the traffic overhead and the convergence time of the algorithms become crucial in larger scenarios, which may entail further research in the field of approximated reinforcement learning techniques.

VI. CONCLUSION

The proposed dynamic solutions seem to achieve a remarkable reduction of the QoE error with respect to the static policy. Such reduction is due to the fact that in the former case the dynamic CoS selection enables per-Application closed-loop QoE differentiation, whereas the latter case has to rely on a network open-loop CoS differentiation mechanisms. The potential impact of the presented approach is huge considering that is leads to a win-win (operators-customers) strategy: indeed the operators exploit their resources not to guarantee aseptic service level agreements, but to optimize their customers' quality perception of the provided services.

Acknowledgments

The authors wish to thank all the members of the FI-WARE and PLATINO consortia, and, in particular, S. Canale, A. Pietrabissa and V. Suraci for their valuable contributions; nevertheless, the concepts and the results contained in this paper are not necessarily the ones of the consortia.

References

[1] FI-WARE (Future Internet Core Platform) EU FP7 project, contract n. 285248, www.fi-ware.eu;

[2] PLATINO (Grant Agreement n° PON01_01007).

[3] Jain, R.; Quality of experience. IEEE Multimedia, Volume: 11 , Issue: 1, 2004, pp. 96-95, DOI: 10.1109/MMUL.2004.1261114

[4] Delli Priscoli, F., "A Fully Cognitive Approach for Future Internet", Special Issue on "Future Network Architectures" of "Future Internet", Molecular Diversity Preservation International (MDPI), Vol. 2, January 2010, pp. 16-29.

[5] ITU-T Recommendation P.10/G.100 *New Appendix I - Definition of quality of Experience (QoE),* 2006.

[6] Gottron C., König A., Hollick M., Bergsträßer S., Hildebrandt T. and Steinmetz R. Quality of experience of voice communication in large-scale mobile ad hoc networks.·In Proceedings of the 2nd IFIP conference on Wireless days (WD'09), pp. 248-253, 2009.

[7] S. Jelassi, G. Rubino, H. Melvin, H. Youssef, G. Pujolle, "Quality of Experience of VoIP Service: A Survey of Assessment Approaches and Open Issues", IEEE Communications Surveys & Tutorials, Vol. 14, Issue 2, 2012.

[8] Goudarzi P. and Nezami Ranjbar M.R. Bandwidth allocation for video transmission with differentiated quality of experience over wireless networks. Computers and Electrical Engineering 37(1), 2011.

[9] S. Singh, J.G. Andrews, G. de Veciana, "Interference Shaping for Improved Quality of Experience for Real-Time Video Streaming", IEEE Journal on Selected Areas of Communications, Vol. 30, Issue 7, Pg. 1259 – 1269, 2012.

[10] Fiedler, M.; Hossfeld, T.; Phuoc Tran-Gia; A generic quantitative relationship between quality of experience and quality of service. IEEE Network, Volume: 24 , Issue: 2, 2010, pp. 36-41, DOI: 10.1109/MNET.2010.5430142.

[11] F. Delli Priscoli, M. Iannone, A. Pietrabissa, V. Suraci, "Modelling Quality of Experience in Future Internet Networks", *Future Network & Mobile Summit 2012,* Berlin, July 2012.

[12] Karagiannis T., Papagiannaki K., and Faloutsos M. BLINC: Multilevel traffic classification in the dark. In ACM SIGCOMM, August 2005.

[13] Lee S., Kim H., Barman D., Kim C.-K., Kwon T.T., Choi Y. NeTraMark: A Network Traffic classification benchmark. In ACM SIGCOMM Computer Communication Review, 41(1), 2011.

[14] M. Castrucci, F. Delli Priscoli, A. Pietrabissa, V. Suraci, "A Cognitive Future Internet Architecture", The Future Internet Future Internet Assembly 2011, Springer Berlin/Heidelberg (DE), Lecture Notes in Computer Science, Vol. 6656, May 2011, doi: 10.1007/978-3-642-20898-0_7.

[15] M. Castrucci, M. Cecchi, F. Delli Priscoli, L. Fogliati, P. Garino, V. Suraci, "Key Concepts for the Future Internet Architecture", Future Network & Mobile Summit 2011, Warsaw, 15-17 June 2011

[16] Delli Priscoli, F., Castrucci, M., "A proposal for future internet architecture", 2010 Future Network and Mobile Summit, 16-18 June 2010, Florence,

[17] F. Delli Priscoli, V. Suraci, M. Castrucci, "Cognitive Architecture for the Internet of the Future". 6th international workshop on next generation networking middleware. October 26-30, 2009. ISBN 978-3-930736-14-0. Pag. 105-112.

[18] A. Pietrabissa, "A Reinforcement Learning Approach to Call Admission and Call Dropping Control in Links with Variable Capacity", European Journal of Control (The European Union Control Association, Lavoisier, France), Vol. 17, Issue 1, 2011, pp. 89-103, ISSN 0974-3580, DOI: 10.3166/EJC.17.89-103.

[19] C. Bruni, F. Delli Priscoli, G. Koch, A. Pietrabissa, L. Pimpinella, "Network decomposition and multi-path routing optimal control" , Transactions on Emerging Telecommunications Technologies (John Wiley & Sons, Inc., USA), Vol. 24, pp. 154-165, 2013, published on line 2 May 2012, doi: 10.1002/ett.2536.

[20] Delli Priscoli, F., Isidori, A., "A control-engineering approach to integrated congestion control and scheduling in wireless local area networks", (2005) Control Engineering Practice, 13 (5), pp. 541-558, doi: 10.1016/j.conengprac.2004.04.016.

[21] F. Delli Priscoli, A. Pietrabissa, "Design of a bandwidth-on-demand (BoD) protocol for satellite networks modeled as time-delay systems", Automatica, International Federation of Automatic Control (Elsevier, Great Britain), Vol. 40, Issue 5, pp. 729-741, May, 2004, DOI:10.1016/j.automatica.2003.12.013.

[22] R. Cusani, F. Delli Priscoli, G. Ferrari, M. Torregiani, "A Novel MAC and Scheduling Strategy to Guarantee QoS for the New-Generation Wireless LAN", Special Issue on "Mobile and Wireless Internet: Architecture and Protocols" of IEEE Wireless Communications (IEEE Personal Communications), IEEE's Computer and Vehicular Technology Societies (U.S.A.), Issue 3, June 2002, pp. 46-56.

[23] Francesco Delli Priscoli, "A Control Based Solution for Integrated Dynamic Capacity Assignment, Congestion Control and Scheduling in Wireless Networks", European Journal of Control, European Union Control Association (EUCA), Issue n.2/2010, pp. 169-184.

[24] A. Palo, L. Zuccaro, A. Simeoni, V. Suraci, L. Musto, P. Garino, "A Common Open Interface to Programmatically Control and Supervise Open Networks in the Future Internet", Future Networks & Mobile Summit 2013, Lisbon, July 2013.

[25] A. Pietrabissa, "Admission Control in UMTS Networks based on Approximate Dynamic Programming", European Journal of Control (The European Union Control Association, Lavoisier, France), Vol. 14, N. 1, pp. 62-75 , January 2008, DOI:10.3166/ejc.14.62-75.

[26] A. Pietrabissa, "An Alternative LP Formulation of the Admission Control Problem in Multi-Class Networks", IEEE Transaction on Automatic Control, (IEEE Control System Society, USA), Vol. 53, N. 3, pp. 839-845, April 2008, DOI: 10.1109/TAC.2008.919516.

[27] F. Delli Priscoli, A. Isidori, L. Marconi, "A Dissipativity-based Approach to Output Regulation of Non-Minimum-Phase Systems", System and Control Letters, Elsevier Science Pub., Vol. 58, 2009, pp. 584-591.

[28] F. Delli Priscoli, L. Fogliati, A. Palo, A. Pietrabissa, "Dynamic Class of Service Mapping for Quality of Experience Control in Future Networks", World Telecommunication Congress (WTC), Berlin, June 2014.

[29] R. Sutton, A. Barto, "Reinforcement Learning: An Introduction", IEEE Transactions on Neural Networks, Vol. 9, pp. 1054–1054, 1998, doi:10.1109/TNN.1998.712192.

[30] Watkins, Christopher JCH, and Peter Dayan. "Q-learning." *Machine learning*8.3-4 (1992): 279-292.

[31] Littman, Michael L. "Friend-or-foe Q-learning in general-sum games." *ICML*. Vol. 1. 2001..

[32] Hu, Junling, and Michael P. Wellman. "Multiagent reinforcement learning: theoretical framework and an algorithm." *ICML*. Vol. 98. 1998.

[33] Fiedler, Markus, Tobias Hossfeld, and Phuoc Tran-Gia. "A generic quantitative relationship between quality of experience and quality of service." *Network, IEEE* 24.2 (2010): 36-41.

[34] Oddi, G., Pietrabissa, A. "A distributed multi-path algorithm for wireless ad-hoc networks based on Wardrop routing," *Proc. 21st Mediterranean Conference on Control and Automation (MED 2013)*, June 25-28, 2013, Platanias-Chania, Crete, Greece, pp. 930-935, doi: 10.1109/MED.2013.6608833.

[35] Macone D., Oddi G., Palo A., Suraci V., "A Dynamic Load Balancing Algorithm for Quality of Service and Mobility Management in Next Generation Home Networks", Telecommun Syst (Springer), 2013, DOI 10.1007/s11235-013-9697-y.

[36] Manfredi, S. "A Reliable and Energy Efficient Cooperative Routing Algorithm for Wireless Monitoring Systems", IET Wireless Sensor Systems 2 (2012), 128-135.

A New Image Encryption Scheme Based on Multiple Chaotic Systems in Different Modes of Operation

Mona F. M. Mursi, Hossam Eldin H. Ahmed, Fathi E. Abd El-samie, Ayman H. Abd El-aziem

***Abstract*—** In this paper, we suggest a proposed image encryption scheme based on multiple chaotic systems in different modes of operation using the development of a Hénon chaotic system also using FRFT which is introduced in order to match the requirements of secure image transfer. We use fractional Fourier transform (FRFT) before the encryption, this process can achieve a large degree of randomization and it makes our proposed algorithm more sensitive to any change in key or plain image. The development of the hénon chaotic map increases the available chaotic range of parameter r to be very wide which make using a development of the hénon chaotic map more robust against attacks. The results of security analysis show that the proposed model provides efficient and secure way for real-time image encryption and transmission.

Keywords: Image Encryption, H □non Map, FRFT.

I. INTRODUCTION

IN recent years, the world lives in the age of communications revolution which necessitates multimedia transmission in a secure manner. Encryption is important in transferring images through the communication networks to protect them against reading, alteration of its capacity, adding false information, or concealing part of its contents [1]. Owing to the frequent flow of digital images across the world over the transmission media, it has become essential to secure them from leakages or threatening or brute force attacks. Our proposed encryption scheme is subjected to encrypting by a combination of shuffling the positions and changing the values of image pixels to shuffle the relationship between the cipher image and the plain image. First, the Baker map is used to shuffle the positions of the image pixels in three different modes of operation (CBC, CFB, and OFB) [6]-[7]-[9]. Second, the shuffled image is encrypted by using our proposed Hénon chaotic map 4, and FRFT before the encryption. This process can achieve a large degree of randomization and it makes our proposed algorithm more sensitive to any change in keys or plain image.

Also in section 2 we describe an analysis of different two dimensional chaotic maps. Section 3 presents Fractional Fourier Transform Domain and show why we use it in image encryption. Our proposed algorithm and experimental results are explained in sections 4, 5 respectively. The conclusions of the paper are presented in section 6.

II. TWO DIMENSIONAL CHAOTIC MAPS

We introduce some different chaotic maps and analyze the simulation results by using MATLAB (R 2013b), with processor Core2 Duo (2.16 GHz) and 2 GB RAM on Windows 8.

A. Baker Chaotic Map

The Baker map is described by the following formulas [8]:

$$B(x,y) = \left(2x, \frac{y}{2}\right)\ 0 \le x < 0.5$$
$$B(x,y) = \left(2x-1, \frac{y}{2} + 0.5\right)\ 0.5 < 0.x \le 1 \qquad (1)$$

1) Generalized Baker map: The Baker map can be generalized in the following way [5], Instead of dividing the square into two rectangles of the same size, the square is divided into k vertical rectangles $[F_{i-1}, F_i]$ x [0, 1], $i = 1, \ldots, k$, $F_i = p_1 + \ldots + p_i$, $F_0 = 0$ such that $p_0 + \ldots + p_k = 1$. The lower right corner of the ith rectangle is located at $F_i = p_1 + \ldots + p_k$. The generalized baker map stretches each rectangle horizontally by the factor of 1/pi. At the same time, the rectangle is contracted vertically by the factor of pi. Finally, all rectangles are stacked on top of each other and we can express Baker map as:

This work was supported in part of the PhD thesis at Department of Computer at Shubra Faculty of Engineering, Benha University.

Mona F. M. Mursi is Prof. Department Of Electrical Engineering, Shubra Faculty of Engineering, Benha University, Egypt, monmursi@yahoo.com.

Hossam Eldin H. Ahmed, Department of Electronics Communication Engineering, P. Dean of the Faculty of Electronic Eng. Menouf-32952, Menufiya University, Egypt, hhossamkh@yahoo.com.

Fathi E. Abd El-samie, Assoc. Prof. Department Of Electronics & Communication Engineering Faculty of Electronic Engineering, Menufiya University, Egypt, fathi_sayed@yahoo.com.

Ayman H. Abd El-aziem, PhD. Student, Department Of Electrical Engineering, Shubra Faculty of Engineering, Benha University, Egypt, eng_aymnhassnin@yahoo.com.

$$B(x,y) = (1/p(x - F_i, P_i y + F_i)$$
$$For, (x,y) \in [F_i, F_i + P_i] \times [0,1] \qquad (2)$$

2) *Discretized Baker map:* Since an image is defined on a lattice of finitely many points (pixels), a correspondingly discretized form of the basic map needs to be derived. In particular, the discretized map is required to assign a pixel to another pixel in a bi-ejective manner. Since the discretized map is desired to inherit the properties of the continuous basic map, the discretized map should become increasingly close to the basic map as the number of pixels tends to infinity. This requirement is expressed mathematically by Equation (3) The discretized generalized Baker map will be denoted B (n_1…, n_k), where the sequence of k integers, n_1,…., n_k is chosen such that each integer n_i divides N, and n_1+…..+n_k = N. Denoting Ni = n_1+…..+n_i, the pixel (r, s), with $N_i \leq r < N_i + n_i$, and $0 \leq s < N$ is mapped [8]-[15].

$$B_{(n1,n2,\dots\dots nk)}(r,s) = \begin{bmatrix} \frac{N}{n_i}\,(r - N_i) + s \bmod \left(\frac{N}{n_i}\right), \\ \frac{n_i}{N}\left(s - s \bmod \left(\frac{N}{n_i}\right)\right) + N_i \end{bmatrix} \qquad (3)$$

This formula is based on the previous geometrical considerations. An N×N square is divided into vertical rectangles of height N and width n_i.

B. Hénon Chaotic System

The Hénon map is a discrete-time dynamical system. It is one of the most studied examples of dynamical systems that exhibit chaotic behavior. The Hénon map takes a point (x_i, y_i) in the plane and maps it to a new point. A Hénon Chaotic system which is described as follows [2]-[3]-[4]:

$$x_{i+1} = 1 - r \times x_i^2 + y_i$$
$$x_{i+1} = b \times x_i \qquad i = 0,1,2,3\ldots \qquad (4)$$

Hénon map, which presents a simple 2-D chaotic map with quadratic nonlinearity, depends on two parameters, r and b, which for the canonical Hénon map have values of r = 1.4 and b = 0.3. For the canonical values the Hénon map is chaotic. For other values of r and b the map may be chaotic, intermittent, or converge to a periodic orbit. An overview of the type of behavior of the map at different parameter values may be obtained from its orbit diagram. For the canonical map, an initial point of the plane will either approach a set of points known as the Hénon strange attractor, or diverge to infinity. The Hénon attractor is a fractal, smooth in one direction and a cantor set in another. This map gave a first example of the strange attractor with a fractal structure. Because of its simplicity, the Hénon map easily lends itself to numerical studies. Thus a large amount of computer investigations followed. Nevertheless, the complete picture of all possible bifurcations under the change of the parameters *r* and *b* is far from completeness.

In the Hénon chaotic map, there are two variables are adopted to encrypt the image. The encryption process consists of three steps of operations [18]:

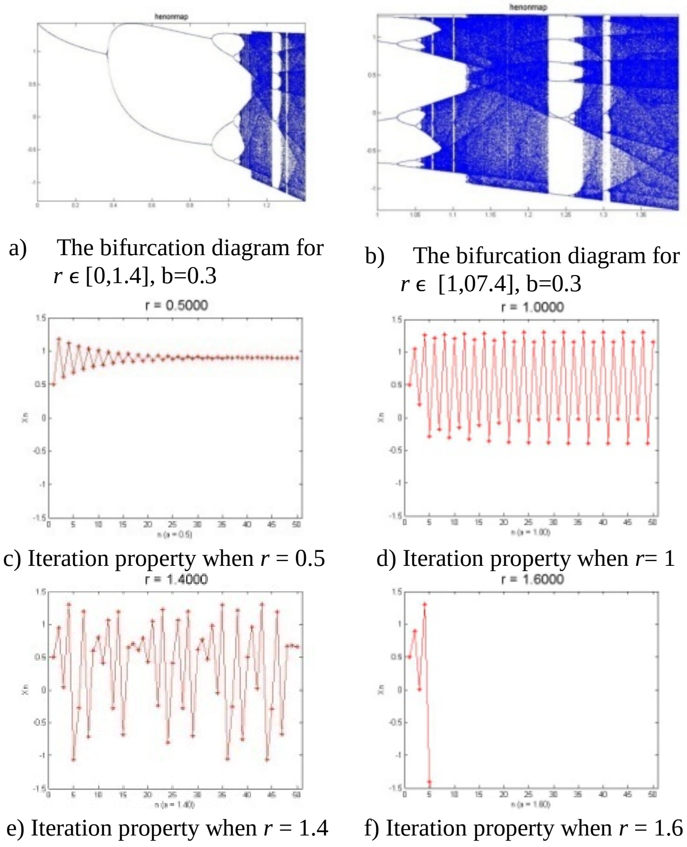
Step1: The Hénon chaotic system is converted into one-dimensional chaotic maps. The one dimensional Hénon chaotic map is defined as:

$$x_{i+2} = 1 - r \times x_{i+1}^2 + b \times x_i \qquad (5)$$

Where *b*= 0.3, *r*∈ [1.07, 1.4]. The parameter *r,* the parameter *b,* initial value x_0 and initial value x_1 may represent the key.

Step2: We adopt a Hénon chaotic map to change the pixel values of the image. First, the Hénon chaotic map is obtained by the equation (5). Then a transform matrix of pixel values is created.

Step3: The exclusive OR operation will be completed bit-by-bit between the transform matrix of pixel values and the values of pixel of the image. The result is the cipher-image. The parameters are selected as b = *0.3*, *r = 1.4*, x_0 = *0.01* and x_1 = *0.02*.



a) The bifurcation diagram for *r* ϵ [0,1.4], b=0.3

b) The bifurcation diagram for *r* ϵ [1,07.4], b=0.3

c) Iteration property when *r* = 0.5

d) Iteration property when *r*= 1

e) Iteration property when *r* = 1.4

f) Iteration property when *r* = 1.6

Fig.1 Analysis of Hénon chaotic map

C. The Proposed Hénon Chaotic Map 1

Our proposed Hénon chaotic map 1 is developed to give a chaotic function which can be used in cryptography application. This Hénon map 1 is expressed as,

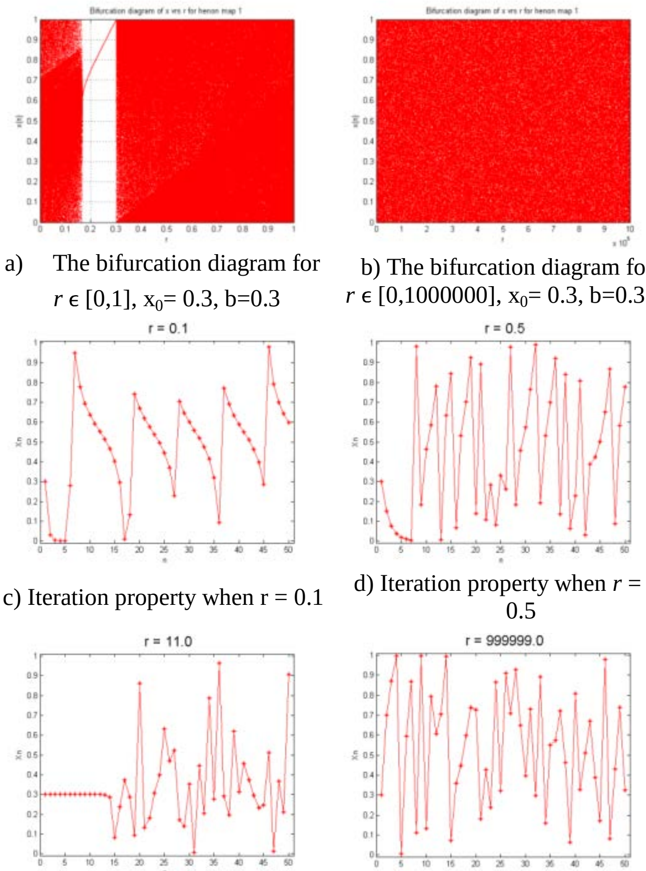
$$x_{i+1} = (r \times x_i + y_i) \bmod 1$$
$$y_{i+1} = 1 - \frac{b}{x_i}\, i = \qquad 0,1,2,3\ldots \qquad (6)$$

The development of the Hénon chaotic map 1 increases the chaotic range of parameter r change from range 1.1to 1.4 to be from 0 to ∞ except from 0.15 to 0.3 and except using by multiple of value 10 as shown in figure 2- a, that will increase the available chaotic value of parameters that can be used in encryption.

The proposed Hénon chaotic map 1 is applied under the following conditions:

- x_n takes a value from interval 0, 1, $x_n \in [0,1]$.
- r is variable and $r \in [0,\infty]$.
- b = 0.3, initial value $x_0 = 0.01$, x_1=0.02.

The simulation is shown in figure 2, it is divided into two parts; one is a bifurcation diagram which draws the relation between x_i for all values of r incremented by 0.001 to show the chaotic behavior of the function, and the second part shows the iteration property of chaotic functions at different values of r to determine which value is suitable for using for encryption as shown in figure 2 (c, d, e, f) shows that we can use any value of r except the range from 0 to 1 and the multiple of value 10.



a) The bifurcation diagram for r ϵ [0,1], x_0= 0.3, b=0.3

b) The bifurcation diagram for r ϵ [0,1000000], x_0= 0.3, b=0.3

c) Iteration property when r = 0.1

d) Iteration property when r = 0.5

e) Iteration property when r = 11

f) Iteration property when r = 999999

Fig.2 Analysis of Hénon chaotic map 1

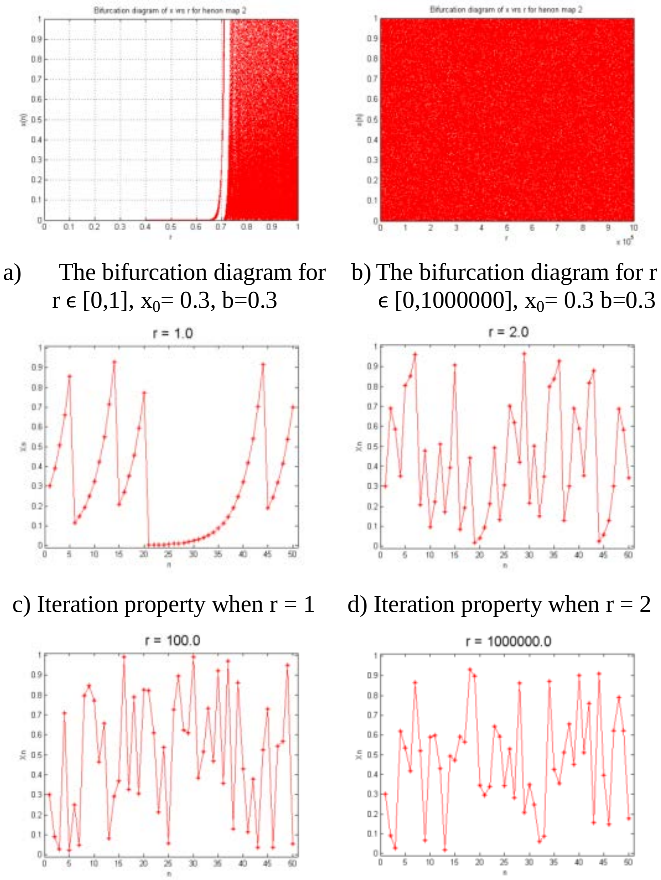
D. The Proposed Hénon Chaotic Map 2

Our proposed Hénon chaotic map 2 is developed to give a chaotic function that can be used in cryptography applications. This Hénon map 2 is expressed as:

$$x_{i+1} = (r \times x_i + y_i) \; mod \; 1$$
$$y_{i+1} = b \times x_i \qquad i = 0,1,2,3\ldots \tag{7}$$

The development of the Hénon chaotic map 2 increases the chaotic range of parameter r to be from 0.7 to ∞ that will increase the available chaotic value of parameter r to be used in encryption.

Our proposed chaotic map is applied under the conditions as proposed Hénon map 1. The proposed Hénon chaotic map 2 overcomes the drawback of the proposed Hénon chaotic map 1 as it can use the multiple values of 10 of variable r shown in figure 3.



a) The bifurcation diagram for r ϵ [0,1], x_0= 0.3, b=0.3

b) The bifurcation diagram for r ϵ [0,1000000], x_0= 0.3 b=0.3

c) Iteration property when r = 1

d) Iteration property when r = 2

e) Iteration property when r = 1000

f) Iteration property when r = 1000000

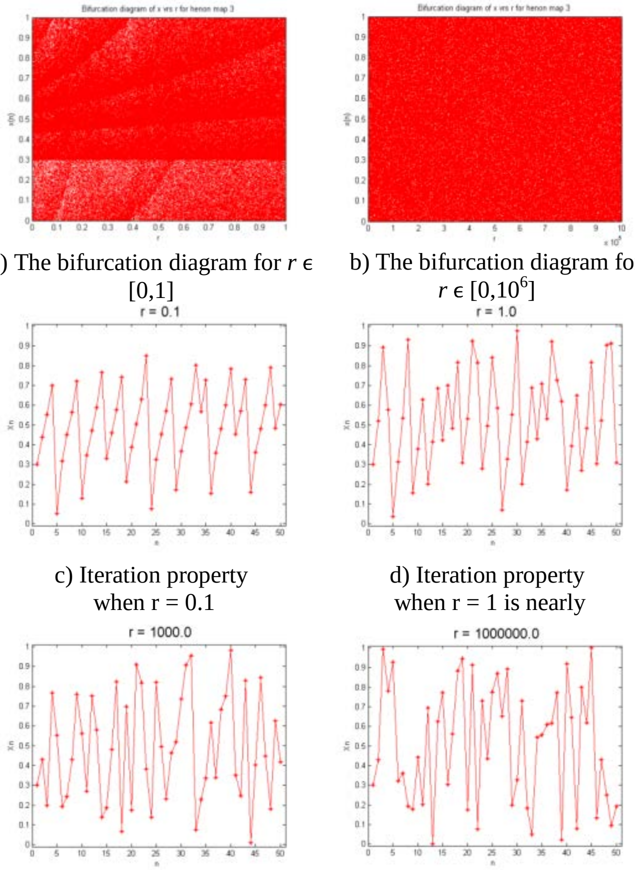
Fig.3 Analysis of Hénon chaotic map 2

E. The Proposed Hénon Chaotic Map 3

Our proposed Hénon chaotic map 3 which is developed to give a chaotic function can be used in cryptography application. This Hénon map 3 is expressed as:

$$x_{i+1} = (r \times x_i + y_i) mod \; 1$$
$$y_{i+1} = \frac{b}{1-x} \qquad i = 0,1,2,3\ldots \tag{8}$$

The development of the Hénon chaotic map 3 increases the chaotic range of parameter r to be from 0 to ∞ that will increase the available chaotic value of parameter r to be used in encryption. It is applied under the conditions as in proposed hénon map 1. The proposed Hénon chaotic map 3 gives a very wide range of variable r for using in encryption and we can use any value of variables r in encryption.



a) The bifurcation diagram for r ϵ [0,1]

b) The bifurcation diagram for r ϵ $[0,10^6]$

c) Iteration property when r = 0.1

d) Iteration property when r = 1 is nearly

e) Iteration property when r = 1000

f) Iteration property when r = 106

Fig. 4 Analysis of Hénon chaotic map 3

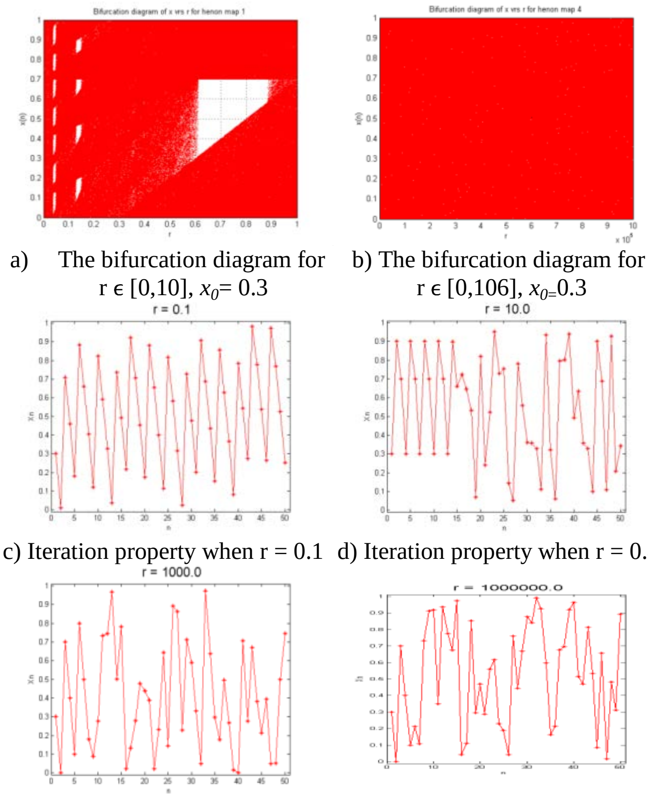
F. *The Proposed Hénon Chaotic Map 4*

Our proposed Hénon chaotic map4 which is developed to give a chaotic function can be used in cryptography application. This Hénon map 4 is expressed as:

$$x_{i+1} = (r \times x_i + y_i) mod\ 1$$
$$y_{i+1} = x_i - b \qquad i = 0,1,2,3\ldots \tag{9}$$

The development of the Hénon chaotic map 4 increases the chaotic range of parameter r to be from 0 to ∞ except value from 0.6 to 0.8 that will increase the available chaotic value of parameter r to be used in encryption. It is applied under the conditions as in proposed Hénon map 1. The simulation is shown in the graph of bifurcation diagram of Hénon chaotic map 4 as shown in the figure 5-a, b, when r ϵ [0,∞]

Also, we use MATLAB to graph the iteration property of Hénon chaotic map4 as shown in the figure 5-c, d, e, f. it becomes a chaotic system without periodicity.



a) The bifurcation diagram for r ϵ [0,10], x_0= 0.3

b) The bifurcation diagram for r ϵ [0,106], $x_{0=}0.3$

c) Iteration property when r = 0.1

d) Iteration property when r = 0.4

e) Iteration property when r = 1000

f) Iteration property when r = 1000000

Fig. 5 Analysis of Hénon chaotic map 4

Finally, we summarize the entire previously proposed Hénon maps in table 1 to compare among them in the available chaotic range of variable r as shown in the table 1.

Table 1 The different Henon Chaotic maps

Chaotic map	**Function chaotic map**	**Range of chaotic for variable r**
Hénon map	$x_{i+1} = 1 - r \times x_i^2 + y_i$ $y_{i+1} = b \times x_i$	[1.07,1.4]
Hénon map 1	$x_{i+1} = (r \times x_i + y_i)\ mod\ 1$ $y_{i+1} = 1 - \frac{b}{x_i}$	[0.3,∞] - multiple of 10
Hénon map 2	$x_{i+1} = (r \times x_i + y_i) mod\ 1$ $y_{i+1} = b \times x_i$	[0.7,∞]
Hénon map 3	$x_{i+1} = (r \times x_i + y_i) mod\ 1$ $y_{i+1} = \frac{b}{1-x}$	[0,∞]
Hénon map 4	$x_{i+1} = (r \times x_i + y_i) mod\ 1$ $y_{i+1} = x_i - b$	[0.6,∞]

III. Fractional Fourier Transform Domain

The Fourier Transform (FT) is one of the most frequently used tools in signal analysis. A generalization of the FT is the FRFT. It has been proposed in [10] and has become a powerful tool for time-varying signal analysis. The FT can be interpreted as a rotation of the signal by an angle of π/2 in the time–frequency plane and represented as an orthogonal signal representation for sinusoidal signals. The FRFT performs a rotation of the signal in the continuous time–frequency plane to any angle and serves as an orthogonal signal representation for the chirp signal. The FRFT is also called a rotational FT or angular FT in some documents. Besides being a generalization of the FT, the FRFT is related to other time-varying signal analysis tools, such as the Wigner distribution, the short-time FT, the wavelet transform.

The chaotic map in the spatial domain has a drawback that keeps the statistical characteristics of the image intact after scrambling. The transform domains provide the ability to transform correlated data patterns into transforming domains to carry the substitution or diffusion process in these domains. This process can achieve a large degree of randomization when returning back to the spatial domain. Chaotic encryption is performed in this transform domain to make use of the characteristics of this domain.

The most intuitive way of defining the FRFT is by generalizing this concept of rotating over an angle that is π/2 in the classical FT case. As like the classical FT corresponds to a rotation in the time-frequency plane over an angle α = 1 π / 2, the FRFT will correspond to a rotation over an arbitrary angle α = aπ / 2 with a ∊ R. This FRFT operator shall be denoted as F multiplied by **a** where F1 = F corresponds to the classical FT operator. The FRFT is defined by means of the transform kernel as [5]:

$$K_\alpha(t,u)\begin{cases}\sqrt{\dfrac{1-jcot\,\alpha}{2\pi}}\times\exp\left(j\,\dfrac{t^2+u^2}{2}\,cot\,\alpha-j\,\dfrac{tu}{sin\,\alpha}\right) if\ \alpha\neq nr\\ \delta(u-t) \qquad if\ \alpha=2n\pi\\ \delta(u+t) \qquad if\ \alpha=(2n+1)\pi\end{cases} \tag{10}$$

The fractional of Fourier transform of a function x, with an angle α , is defined as the function R $^{\alpha}$ =X $^{\alpha}$,

$$X_\alpha(u)=\int_{-\infty}^{\infty}x(t)k_\alpha(t,u)dt \tag{11}$$

One of them is a kernel-based integral transformation of the form:

$$\begin{aligned}f_a(u)&=\ F^a[\,f(x)]\\&=C_\alpha\int f\,(x)\exp\left[i\,\pi\,j\,\frac{t^2+u^2}{tan\,\alpha}\,cot\,\alpha\right.\\&\qquad\left.-\ 2i\pi\,\frac{ux}{sin\,\alpha}\right]dx\end{aligned} \tag{12}$$

Where $\alpha=\frac{a\pi}{2}$, and $C_\alpha=\frac{\exp\left[-\mathrm{i}(\frac{\pi.\mathrm{sign}(\sin\alpha)}{4}-\frac{\alpha}{2})\right]}{|\sin\alpha|^{1/2}}$

IV. A PROPOSED IMAGE ENCRYPTION ALGORITHM BASED ON MULTIPLE CHAOTIC SYSTEMS IN DIFFERENT MODES OF OPERATIONS

In this section, the proposed image encryption scheme is based on using a multiple of chaotic maps in different modes of operation in the fractional Fourier domain as shown in figure 6. The proposed algorithm is divided into two parts: first applying the FRFT to the original image; the second part combines the confusion with diffusion. The confusion algorithm using 2-D chaotic baker map scrambling in three different modes of operations; Cipher Block Chaining (CBC), Cipher Feedback (CFB), Output Feedback (OFB), in which the Initialization Vector (IV) works as the main key. The diffusion applied on the development of Hénon chaotic system. The diffusion algorithms are using one of the proposed hénon chaotic maps as hénon map 4. There are many researches that study the effect of modes of operation on the security as [12-13-18], but in this paper we introduce it by new proposed algorithm.

The proposed algorithm cryptosystem has high security performance as it fulfills the classic Shannon requirements of confusion and diffusion [11]. We examine its implementation for digital images along with its detailed security analysis to study the effect of modes of operation on the performance of chaotic cryptosystem implemented in the FRFT domain. Besides that, applying the development of the hénon chaotic map to increase the security and make the relation between image and encrypted image more complex. We examine its implementation for digital images along with its detailed security analysis. The security analysis includes the statistical analysis, the histogram analysis, the correlation coefficient metric, the maximum deviation metric, the irregular deviation metric and the processing time.

The proposed algorithm steps are described in the following steps:

1. Applying the fractional Fourier transform to the original image.
2. The transformed image is encrypted using the chaotic baker map in three different modes of operation for confusion.
3. The shuffled transformed image is applied to the second stage of chaotic encryption for diffusion (one of the proposed hénon maps). Hence, we obtain the ciphered image.

We have implemented the baker chaotic map with five different values of W (the block size) as follows:

1. $W_1$is 128×128 pixels, and the initialization vector and (IV_1) was a section of the encrypted Cameraman image.
2. W_2 is 64×64 pixels, and $IV_2$was a section of the encrypted Cameraman image.
3. W_3 is 32×32 pixels, and IV_3 was a section of the encrypted Cameraman image.
4. W_4 is 16×16 pixels, and the initialization (IV_4) was a section of the encrypted Cameraman image.
5. $W_5$is 8×8 pixels, and $IV_5$was a section of the encrypted Cameraman image.

The diffusion process of a proposed hénon map 4 scheme, consists of three steps of operation as steps of hénon chaotic map. The diagram of the combination of the confusion and diffusion algorithm with FRFT to produce cipher image is shown in figure 6.

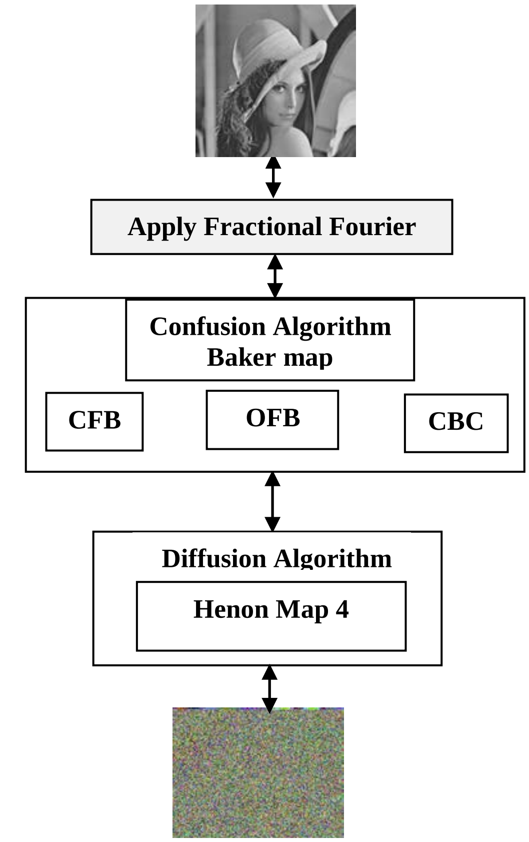
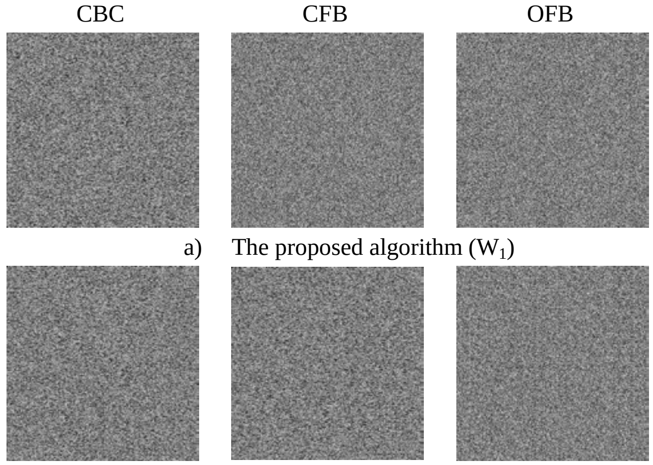


Fig. 6 The block diagram of the proposed Hybrid cryptosystem.

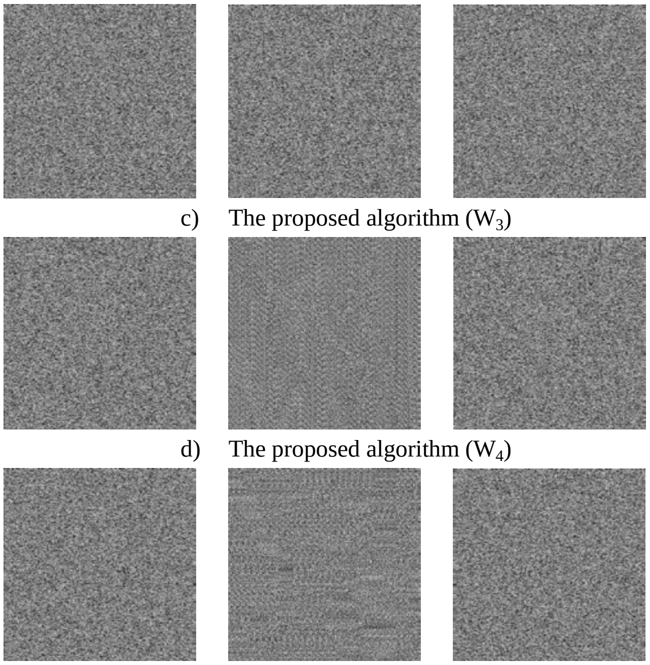
V. ENCRYPTION EVALUATION METRICS AND EXPERIMENTAL RESULTS

One of the important factors in examining the encrypted image is the visual inspection. It is clear there are any details of using our encrypted image using encryption algorithm, but it is not sufficient to depend on the visual inspection only. So, other metrics are considered to evaluate the degree of encryption quantitatively [9]. The encrypted images of Lena using the proposed algorithm are shown in figure 7.



a) The proposed algorithm (W_1)

b) The proposed algorithm (W_2)



c) The proposed algorithm (W_3)

d) The proposed algorithm (W_4)

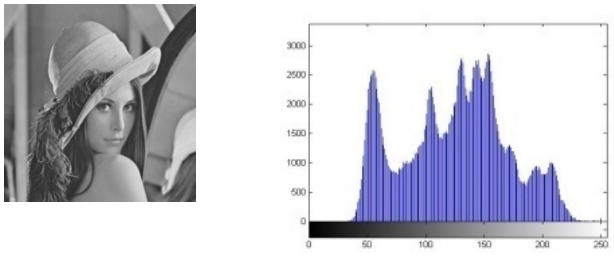
e) The proposed algorithm (W_5)

Fig.7 The encrypted image using proposed algorithm in three modes of operation with different block sizes (W_1,.. , W_5).

A. *Statistical Analysis.*

To examine the quality of encryption and the stability via statistical attacks, the histogram is calculated for all images, correlation coefficient (CC) between original image and cipher-image, maximum deviation factor (MD), and irregular deviation factor (ID).

1) *Histogram analysis:* The original image (lena.bmp) with the size 512×512 pixels is shown in Figure 8(a) and the histogram of the original-image is shown in Figure 8(b). Figure 9 illustrates the histogram of the encrypted images using our proposed algorithm with different modes of operation. As we can see, the histograms of the encrypted image using our proposed algorithm in different modes of operation with different block sizes are fairly uniform and are significantly different from that of the original image. This is because the diffusion caused by the effect of modes of operation beside the diffusion caused by applying our proposed h □non chaotic map 4. So, modes of operation beside h □non chaotic map 4 and using chaotic cryptosystem applied in FRFT domain improve the histogram of the encrypted images.



(a) The original image. (b) The histogram of the original image.

Fig. 8 The histogram of the original image

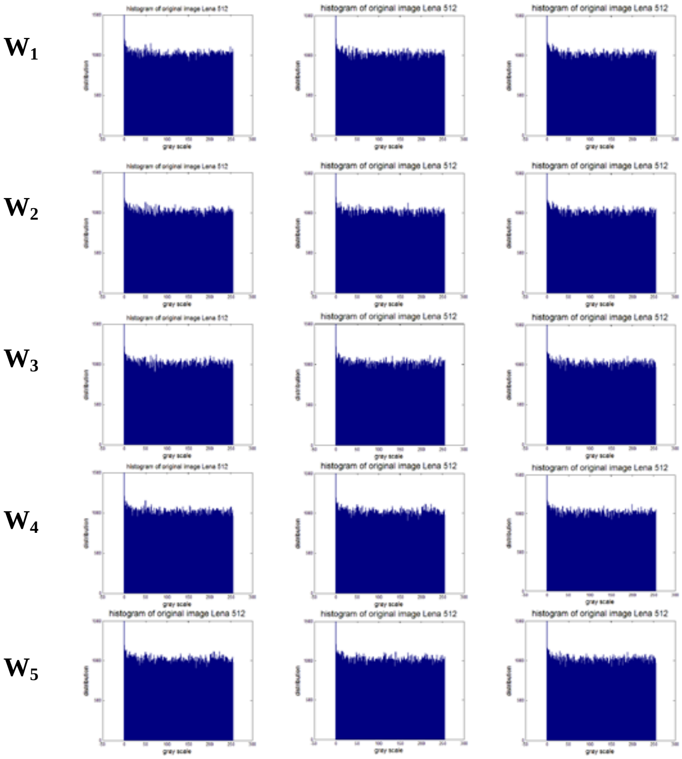


Fig. 9 The Histogram of encrypted image using the chaotic baker map in different modes of operation with different block sizes (W_1, . . . , W_5).

2) *Correlation coefficient analysis:* The correlation coefficient equals one if they are highly dependent, i.e. the encrypted image is the same as the original image and the encryption process failed in hiding the details of the original image. If the correlation coefficient equals zero, then the original image and its encryption are totally different. So, success of the encryption process means smaller values of the CC. The CC is measured by the following equation [9-16-17]:

$$r_{xy} = \frac{cov\ (x, y)}{\sqrt{D\ (x)}\sqrt{D\ (y)}} \tag{13}$$

Where x and y are the gray-scale values of two pixels at the same indices in the plain image and cipherimage. In numerical computations, the following discrete formulas can be used:

$$E\ (x) = \frac{1}{L}\sum_{i=1}^{L} x_i \tag{14}$$

$$D(x) = \frac{1}{L}\sum_{i=1}^{L}(x_i - E(x))^2 \tag{15}$$

$$cov(x, y) = \frac{1}{L}\sum_{i=1}^{L}(x_i - E(x))(x_i - \ (y)) \tag{16}$$

Where L is the number of pixels involved in the calculation

Table 2 illustrates the correlation coefficient between the original image and the encrypted image using our proposed algorithm with different modes of operation and different block sizes. It is clear that the correlation coefficient decrease as the block size increase. This is due to the fact that the number of XOR operations used in any mode of operation to decrease as the block size increase.

Table 2 The correlation coefficient between the original image and encrypted images using the proposed algorithm in three modes of operation CBC, CFB, OFB with different block sizes.

Mode of operation	**Block size**				
	W_1	**W_2**	**W_3**	**W_4**	**W_5**
CBC	0.0	0.0002	0.0008	0.0044	0.0041
CFB	0.0	0.0018	0.0	0.0041	0.0011
OFB	0.0	0.0006	0.0011	0.0034	0.0011

3) *Maximum deviation analysis:* The maximum deviation measures the quality of encryption in terms of how it maximizes the deviation between the original and the encrypted images [9]-[17]. Table 3 illustrates the maximum deviation measuring factor [9]-[16] of the encrypted images. As we can see, the CBC mode with W3 achieved better results than the OFB mode with W2, and the CFB mode with W5 makes the worst results between all modes, but CBC with W3 has the best result.

Table 3 The maximum deviation metric of the proposed algorithm in three modes of operation with different block sizes.

Mode of operation	**Block size**				
	W_1	**W_2**	**W_3**	**W_4**	**W_5**
CBC	187898	187779	188346	187220	187850
CFB	187973	187010	187320	1878500	1867200
OFB	188285	188285	188100	187940	188100

4) *Irregular deviation analysis:* This analysis is based on how much the deviation caused by encryption (on the encrypted image) is irregular. This method can be summarized in some of steps to obtain irregular deviation (ID) [9].The irregular deviation metric can be used alone to test the quality of encryption in the field of image encryption. So, if this factor agrees with other metrics, it will be a good judge, otherwise the final decision on measuring the quality of the encryption algorithms will be on the irregular deviation on this test. From the results shown in Table 4, the CFB with W_3 has better results than the CFB mode with W_1. The OFB mode with W_4, W_5 has the worst result.

Table 4 The irregular deviation metric of the proposed algorithm in three modes of operation with different block sizes

Mode of operation	**Block size**				
	W_1	**W_2**	**W_3**	**W_4**	**W_5**
CBC	180464	180602	180676	181096	181114
CFB	180378	181084	180296	181114	181286
OFB	180826	180826	180988	180962	180988

B. NPCR and UACI Analysis.

To evaluate the variations between the original image and the decrypted images, there are two additional tests: NPCR (Number of Pixels Change Rate) and UACI (Unified Average Changing Intensity). NPCR and UACI are performed as follows [10]:

Table 5 The UACI between encrypted image and original image in different modes with different block sizes.

Mode of operation	Block size				
	W_1	W_2	W_3	W_4	W_5
CBC	28.6269	28.6334	28.6053	28.5494	28.5549
CFB	28.6140	28.5485	28.6469	28.5549	28.5595
OFB	28.6260	28.6262	28.5988	28.5495	28.5988

Table 6 The NPCR between encrypted image and original image in different modes with different block sizes.

Mode of operation	Block size				
	W_1	W_2	W_3	W_4	W_5
CBC	99.61	99.61	99.62	99.60	99.63
CFB	99.61	99.59	99.61	99.63	99.61
OFB	99.61	99.61	99.63	99.63	99.63

C. Time Analysis

In this analysis, the processing time has also been tested here. First, it is defined as the times required to encrypt/decrypt data. The smaller the processing time, the higher the speed of encryption is. We have tested the influence of the block size on the processing time of the proposed cipher implemented in CBC, CFB, and OFB modes. Table 7 shows the processing time for the encryption process for the algorithm in different modes of operation with different block sizes. From this table we can conclude that the processing time doesn't depend on the mode of operation used, and the processing time increases with the decrease of the block size. This is due to the fact that the number of XOR operations used in any mode of operation increases with the decrease of the block size.

Table7. The processing time of the encryption process versus block size in sec.

Mode of operation	Block size				
	W_1	W_2	W_3	W_4	W_5
CBC	2.63	2.63	3.15	5.81	15.41
CFB	2.89	2.53	3.05	5.82	15.51
OFB	2.96	2.78	3.43	5.76	15.04

D. key space analysis

A good encryption scheme should resist most kinds of known attacks. It should be sensitive to the secret keys, and the key space should be large enough to make brute-force attacks infeasible. In our proposed algorithm, the key space analysis and tests are summarized in the following sections.

1) Exhaustive key search: For a secure image cryptosystem, the key space should be large enough to make the brute-force attack infeasible. An exhaustive key search will take 2^k operations to succeed, where k is the key size in bits. An attacker simply tries all keys, and this will be very exhaustive. For chaotic maps encryption, the key is dependent on the width (or height) of the image to be encrypted. This is due to the scrambling phenomenon of the chaotic map. For the 512x512 Lena image, the number of possible keys $=10^{128}$. Thus, in this case the computations will require:

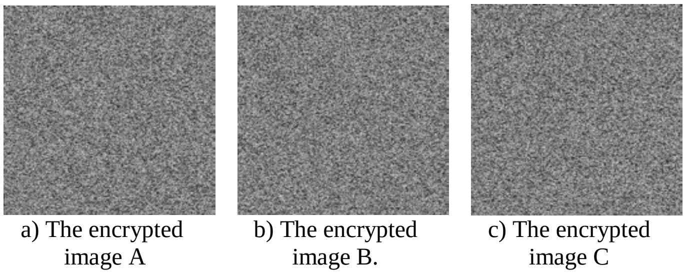
$$\frac{10^{128}}{1 \times 10^{9} \times 60 \times 60 \times 24 \times 365} = 4.3675 \times 10^{111} \qquad (17)$$

This is practically infeasible. Beside that we have the parameter r, b, x0, x1 of our proposed hénon chaotic map are another key and the angle of FRFT, all these parameters are key of our proposed algorithm.

2) Key sensitivity test: Large key sensitivity is required by secure image cryptosystems, which means that the cipher image cannot be decrypted correctly if there is only a slight difference between encryption and decryption keys. For the proposed algorithm, we test the sensitivity of the key by two methods, fist we make small changes in one parameter of key (r, b, x_0 ,x_1,) of the our proposed hénon map4. We can also test the sensitivity by a change in angle of FRFT by changing the value **a**; change key which using an encryption we can perform the following steps:

1. The original image is encrypted using the secret key using b=0.4, r=11 x0 = 0.01, x1=0.02 and the encrypted image using can be shown in figure 10-a.
2. The same image is encrypted by making a slight modification in the secret key using *r*=11.1 the encrypted image can be shown in figure 10-b.
3. Again, the same image is decrypted by making another slight modification in the secret key using *r*= 11.2, the encrypted image can be shown in figure 10-c.

Finally, the three encrypted images A, B, and C are compared.



a) The encrypted image A

b) The encrypted image B.

c) The encrypted image C

Fig.10 Applying the proposed algorithm using different keys

It is not easy to compare the encrypted images by simply observing them. Thus, for comparison, we can calculate the correlation coefficients between the original image and the three encrypted images. The lower the correlation coefficient, the higher key sensitivity is.

Table 8 The correlation coefficients between the three encrypted images A, B, and C

Image 1	Image 2	C.C
Encrypted image A	Encrypted image B	-0.0006
Encrypted image B	Encrypted image C	0.0014
Encrypted image A	Encrypted image C	0.0017

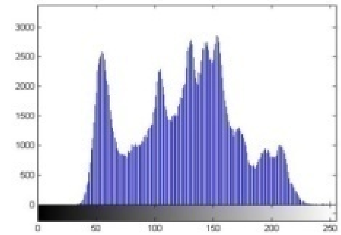
From this table we can see that the correlation coefficients for the three cases are low thus proving a high key sensitivity and the three different images are completely different.

Second method for test sensitivity of our proposed algorithm can be done as follows: we test the sensitivity of the key by making a small change in the constant r,b, xo and x1 of the our proposed hénon map4. We can also test sensitivity by changing the angle of FRFT by changing the value a1; changing the key which is used in encryption, we can perform the following steps:

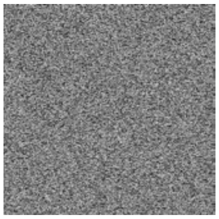
1. The original image is encrypted using the secret key using, r=11, b=0. 4, x0 = 0.01, x1=0. 02, the original image is shown in figure 11-a and the decrypted image using the same key shown in figure 11-b it is clear that the decrypted as original image and its histogram has been illustrated in figure.11-c.
2. The same image is decrypted by making a slight modification in the secret key using r=11.000000000000001, such that r is changed a little (10^{-15}), as shown in figure 3.11-b the decrypted image is completely different than the original image The original image is encrypted using the secret key using b=0.4, r=11 x0 = 0.01, x1=0.02antd the encrypted image using the same key as shown in figure 11-d it is clear that the decrypted as the encrypted image and its histogram has been illustrated in figure.11-e.
3. Again, the same image is decrypted by making another slight modification in the secret key using *x1* = 0.0100000000001, such that *x1* is changed a little (10^{-13}). As shown in figure 3.11-f the decrypted image is completely different than the original image and its histogram has been illustrated in figure 11-g.

a) The original image

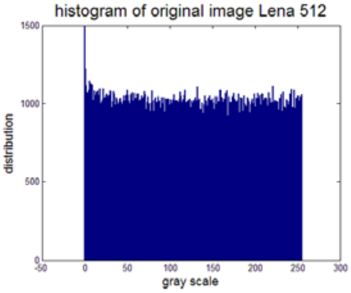
b) decrypted image at *r*=11, *b*=0.4, *x0*=0.01, *x1*=0.02



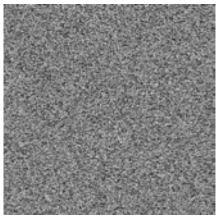
c) Histogram of the decrypted image.



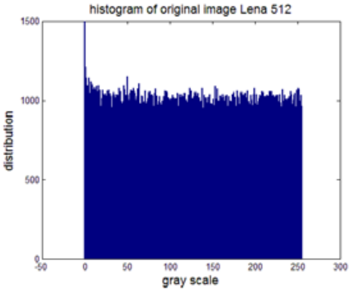
d) decrypted image (at *b*=0.4, *x0*=0.01, *x1*=0.02 *r*=11.000000 000000001,



e) Histogram of the decrypted image.



f) decrypted image at r=*11*, *b*=0.4, *x1*=0.02 *x0*=0.0100000000001,



g) Histogram of the decrypted image.

Fig.11 Applying the proposed algorithm using different keys

VI. CONCLUSION

This paper focuses on two main parts: first part presents the development of hénon chaotic map. The development of hénon chaotic map increases the available chaotic range of parameter to be very wide, which makes the using of development of the hénon chaotic map more robust against attacks because we have a wide range of r is available to use compared to the limited value of variables are in the standard hénon chaotic map.

The second part introduced, a proposed encryption scheme based on multiple of chaotic system using different modes of operation. The proposed is combining confusion algorithm using baker chaotic map in the three modes of operations CBC, CFB, and OFB with a diffusion scheme for encrypting the images by changing the pixels values of the image by using proposed hénon map 4. All of these procedures for encryption are used in FRFT domain. The experimental results and analysis show that the proposed cryptosystem is the best and has high security such that,

1. The proposed scheme has a large enough key space to resist most kinds of brute force attacks.
2. It is very sensitive to all members of the secret keys.
3. Their result of NPCR and UACI tests gives good results.
4. The proposed scheme is suitable to provide an efficient and secure way for image encryption and it executes in a short time for encryption, when using large block size in any mode of operation.
5. The modes of operation can improve the performance of chaotic cryptosystem and our proposed algorithm with CBC give the best result.

REFERENCES

[1] N. k. Pareek, Vinod Patidar, K. K. Sud.” Image Encryption Using Chaotic Logistic Map” Image and Vision Computing 24, P.P(926-934) 2006.

[2] E. Petrisor.. Entry and exit sets in the dynamics of area preserving Hénon map. Chaos, Solitons and Fractals, pp. 651–658, Oct. 2003.

[3] L. Guo-hui, Z. Shi-ping, X. De-ming, L. Jian-wen." An Intermittent Linear Feedback Method for Controlling Hénon-

like Attractor". Journal of Applied Sciences, pp. 288–290, Dec 2001.

[4] Chen Wei-bin, Zhang Xin. "Image Encryption Algorithm Based on Hénon Chaotic "System.978-1-4244-3986-7/09/$25.00 © IEEE 2009.

[5] L. B. Almeida, "The fractional Fourier transform and time-frequency representations," *IEEE Trans. Signal Process.*, vol. 42, no. 11, pp.3084–3091, Nov. 1996.

[6] Swiss encryption technology, "MediCrypt, Modes of operation", http://www.mediacrypt.com/pdf/MC modes1204pdf, pp. 1-4.

[7] Clifiord Bergman, Encryption modes, pp. 1-18, Lecture 16, Feb. 2005,

[8] Jiri Fridrich" Symmetric Ciphers Based on Two-Dimensional Chaotic Maps" International Journal of Bifurcation and Chaos, Vol. 8, No. 6 , 1259-1284, 1998.

[9] Nawal El-Fishawy, Osama M. Abu Zaid Quality of Encryption Measurement of Bitmap Images with RC6, MRC6, and Rijndael Block Cipher Algorithms. International Journal of Network Security, 5(3) : 241–251, Nov. 2007.

[10] B. M. Hennelly and J. T. Sheridan, "Image encryption based on the fractional Fourier transform," *Proc. SPIE*, vol. 5202, pp. 76–87, 2003.

[11] C.E. Shannon, A Mathematical Theory of Communication, Bell Sys. Tech. J. 27:379 {423, 623 {656,} 1949.

[12] Heba M. Elhosany, Hossam E. Hossin, Alaa M. Abbas, Hassan B. Kazemian, Osama S. Faragallah, Sayed M. El-Rabaie, Fathi E. Abd El-Samie, "Chaotic encryption of images in the Fractional Fourier Transform domain using different modes of operation," Signal, Image and Video Processing, ISSN 1863-1703, Vol 7, Issue 3, May 2013.

[13] I. F. Elashry, O. S. Faraga llah, A. M. Abbas, El-Sayed M. El-Rabaieand, F. E. Abd El-Samie, "Chaotic Image Encryption with Different Modes of Operation" ,IET Image Processing 2009.

[14] Osama S. Faragallah , E. M. Nigm, Nawal A. El-Fishawy, Osama M. Abu Zaid "A Proposed Encryption Scheme based on Hénon Chaotic System (PESH) for Image Security "International Journal of Computer Applications (0975 – 8887) Volume 61– No.5, January 2013 .

[15] J. Fridrich. Image encryption based on chaotic maps. In Proc. IEEE Int. Conference on systems, Man and Cybernetics, volume 2, pages 1105–1110, 1997.

[16] H. Elkamchouchi and M. A. Makar, "Measuring encryption quality of Bitmap images encrypted with Rijndael and KAMKAR block ciphers," in Proceedings Twenty second National Radio Science Conference (NRSC 2005), pp. C11, Cairo, Egypt, Mar. 15, 17, 2005.

[17] I. Ziedan, M. Fouad, and D. H. Salem, "Application of Data encryption standard to bitmap and JPEG images," in Proceedings Twentieth National Radio Science Conference (NRSC 2003), pp. C16, Egypt, Mar. 2003.

[18] Osama S. Faragallah, E. M. Nigm, Nawal A. El-Fishawy, Osama M. Abu Zaid Osama M. Abu Zaid "A Proposed Encryption Scheme based on Hénon Chaotic System (PESH) for Image Security " International Journal of Computer Applications (0975 – 8887) Volume 61– No.5, January 2013

Collaborative and Integrated Designing of Intelligent Sustainable Buildings

Luminita Popa, Simona Sofia Duicu

Abstract— The purpose of this paper is to describe a method of collaboration dedicated to a holistic and integrated design of a intelligent building in a sustainable virtual environment. Given the increasing complexity of durable built spaces, quality, cost, time and environment constraints, it is necessary to review the principles of classic design and product development. Significant improvements in the process design can be achieved by integrating CAD and CAE models to work in perfect harmony liaising interdisciplinary teams manifesting the spirit of collaboration in achieving the original construction. Technologies and intelligent buildings design processes (built durable) use today, have varying degrees of success. It is commonly known that during design process, engineers are using one or more computer-aided graphics programs.

Keywords— Intelligent Sustainable Buildings; Collaborative and Integrated Design; Intelligent Energy Efficiency

I. Buildings Holistic and Intelligent Design

The basic mission of a multidisciplinary research for a sustainable built environment is the holistic intelligent design. This is the crucial growth accelerator for intelligent sustainable buildings The solution to achieve a truly energy-efficient building is interoperability, which is developed when the associated components with a variety of systems, regardless of manufacturer, "talk to each other" and work together smoothly to full operational efficiency.

However, over the last decade, the prevalence of proprietary protocols failed to provide the desired opening and slowed market development. Open architecture is a key factor for the adoptions of intelligent building models, in turn, are essential for connecting and creating smart grid.

A smart grid had borrowed the concepts from the Internet, such as connecting products and systems to create a network of components that communicate with one another in real time (Fig.1).

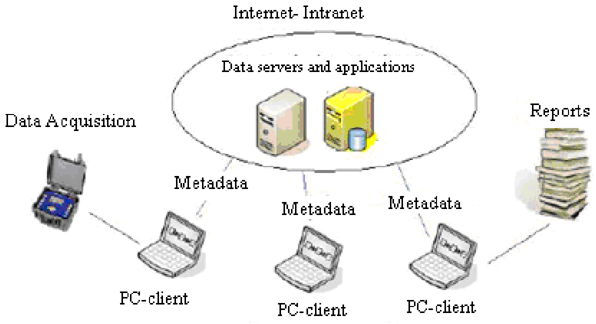


Fig. 1 Collaborative engineering within global internet

An open architecture establishes communication protocols that enable building control systems from different vendors to exchange information, to synchronize equipment and get optimal functioning of the intelligent sustainable building.

While in the past all protocols were proprietary, nowadays open protocol becomes an optional technology, providing a high level of energy and efficiency management. Open architecture improves control and simplifies intelligent sustainable buildings design

Traditionally, energy downstream devices, such as: HVAC (Heating, Ventilation&Air Conditioning), lighting and access control, operate separately within the building.

When systems from different manufacturers are not compatible, building owners are faced with a difficult task if they want to improve the energy performance of their buildings.

When the equipments from different suppliers must exchange information, are necessary interfaces (gateways). This creates the need for custom software and hardware, with considerable time and financial investment for installation and maintenance.

The possible synergies by integrating isolated systems increase the number of strategies available for managing various products and reduce energy consumption. End users can choose the best options available in terms of investment and functionality, and can benefit of their control systems to make a building more energy efficient, cheaper and greener.

Other benefits include: a single user interface; maintenance and training low cost for obtaining a future system with more flexibility to plan with the increase in time the organization.

Systems based on open communication standards and add to the building a real value.

L. P.Author is with the, Transilvania University of Brasov, Department of Automation and Information Technology , CO 500174 România (corresponding author to provide phone: +4-0721-243-440; fax: +4-0268-418836; e-mail: mluminita2001@ yahoo.com)

S. S. D Author is with the Transilvania University of Brasov , Department of Engineering and Industrial Management, CO 500174 România (corresponding author to provide phone: +4-0722-410-515; fax: +4-0268-418836; e-mail: simonaduicu@unitbv.com)

II. Intelligent building integrated design

The concept of intelligent design is the result of a shared vision that tends to merge leading achievements in areas that involve the design and construction of buildings, areas usually considered separately. The most important of these areas include: external and internal architecture, construction technology in general, along with lighting, heating, ventilation, communication, computer, inland transport, security, ergonomic factors, etc.

The intelligent design features of systems are as following:

- cooperation of all existing issues in the design process from conceptual design stage itself;
- aggregation and cooperation of existing knowledge in the design process from conceptual design stage itself;
- integration and cooperation of description and data representation models that must be unique and common in all phases of design, which means not only changing the interface;
- systems intelligence, which includes intelligent support for designer, intelligent interfaces, intelligent functions (knowledge base, inference capability that can understand the intentions of the designer, to detect errors, to suggest alternatives, to answer questions, etc.); the most important of these is the detection of errors that reduce design time and cost;
- reducing time error detection, increasing the accuracy of analysis or synthesis, reducing costs, particularly the maintenance costs.

Within this approach, project participants work is not sequential, but has a certain roundness, the date of return (to improve conception building) decisions of the previous stages.

In the traditional approach, unconcerned with the building energy performance, a building designers are generally conduct their activities in a sequential manner:
- Beginning of the process, which is decisive for the final performance of the building, is made by architect without an assessment of the energy efficiency
- Structure engineer makes their share starting from what the architect designed.
- Engineers for building utilities/services (ventilation, heating / cooling, hot/cold water, electricity) bring their own contributions to the project, seeking to accommodate to the architect and structural engineers works. In this late phase is already too expensive to increase the energy performance of the building, other than minor changes to the architecture and strength.

Designing utilities / services is already influenced by the architecture and data structure of the building.

Leeway for increasing energy efficiency of the building is limited, the energy calculation being more confirming than generating strategic decisions for building design. Right from the start, the architect works closely with at least one energy consultant. They are working with utilities engineer in an iterative mode to evaluate several options and to select the most optimal energy. An example of virtual cooperation between building and construction department, HVAC design department and electrical design department is shown in Fig. 2.

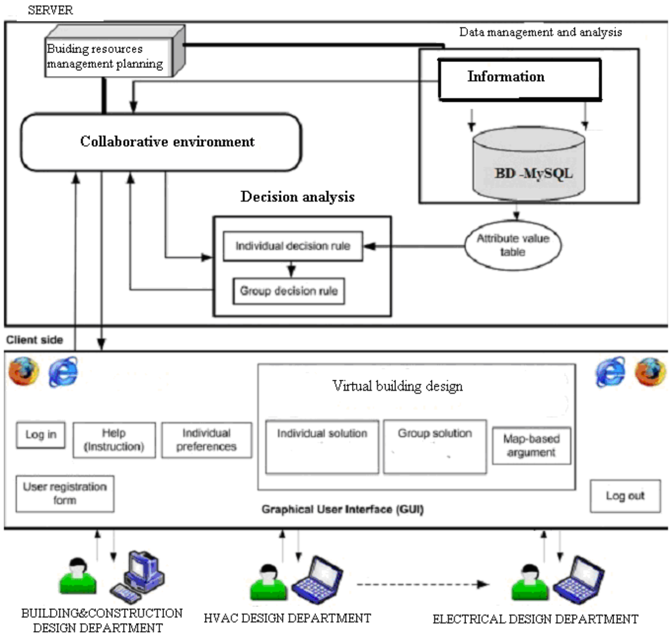


Fig.2. Virtual cooperation between building and construction department, HVAC design department and electrical design department

Starting from an initial proposal (eg. a given shape and orientation of the building, a particular shape and arrangement of windows, a specific structure of the walls and floors and a certain way of combining them, some basic ideas about building utilities /services, especially heating/cooling and ventilation installations) is doing the calculation of the annual energy required to operate the building.

Calculation returns to the starting point by performing some modifications, alleged beneficial and the energy is recalculated. The procedure is repeated for several options in order to select the optimal case

This iterative return for decisions modification of a previous stage is made in order to increase the energy performance of the building and can be done in several phases of the design process

But it is generally recognized that decisions even in the early stages of conceptual design of the building, have the most significant impact on the building final energy performance.

In order to obtain maximum performance it is necessary to conduct an integrated design process in which the architect works from the very beginning together with a resistance engineer and a design engineer specializing in energy and climate control systems to optimize energy efficiency and building sustainability .

Thus, the traditional working teams are converting in virtual teams and operate to/from away from Headquarters in the so-called "workgroups sites".

Virtual teams or workgroups are using tools and information and communication technologies (ICT) equipments.

III. Sustainable Intelligent Building Systems

"Sustainable development meets the requirements of the present without compromising the ability of future generations to meet their own needs" - United Nations Commission on Environment and Development

Sustainability includes the following elements:

- Minimize the actions that degrade the planet's ecosystems and living resources.
- Actions that aim to restore and sustain these systems and resources.

The Intelligent Energy Efficiency Building Solution project proposes to enable real-time measurement, monitoring and management of building systems through the development of new algorithms, analytics of real-time external events and building system state, and implementation of control mechanisms to better optimize energy so consumption can be reduced (Fig.3).

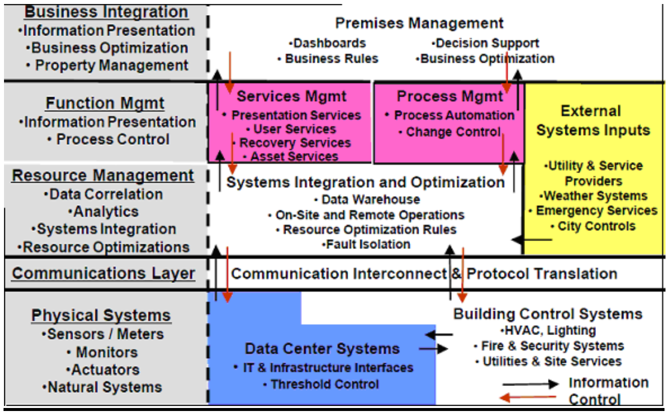


Fig.3. Intelligent Building High Level Architecture (adapted from [1]

The concept of "Intelligent Building" refers to the interconnecting possibility of modern communication technologies which can make available to the customer requirements present in every home.

With these intelligent systems, it is possible to control electrical devices and security systems from any location on the globe, having available a phone or a computer.

The intelligent building enables owners and managers to optimize the potential benefits and savings when the equipments associated with various building systems (eg. HVAC, fire prevention solutions and security) and/or the buildings are managed integrated into a single point, by using a centralized automation and control system. The user will be guided to perform any operation, be it a control lights, air conditioning or to enable or disable the security system. In fuctional terms, the areas of interest of these intelligent systems are: automation, connectivity, entertainment and access to the information.

An open architecture establishes communication protocols that enable building control systems from different vendors to exchange information, to synchronize equipment and get optimal energy functioning of the building. Within intelligent buildings, intelligent management and control solutions include building management and energy convergence with security, life safety and fire safety, communication, IT equipments rooms, automation machines and specific applications to create a highly adaptable building , durable and economical.

In conclusion, the energy-efficient design, really comes from the growing need to manage the whole building as an integrated system.

IV. Automation as a Part of Intelligent Building High Level Architecture

Automation concept involves the development of devices that do not require the intervention of the human operator to realize the purposes for which they were created.

In general, when we are talking about automation, people come in head images with a robot hand which is carrying things on assembly line (eg. to fix a door to a car, inserts some integrated circuits on a printed circuit board, etc.).

In reality we are thinkig about the devices which know what actions to take in case of special events.

The simplest example is the ability to program the house turning on lights with dusk while we are out of town. There are already vacuums that "walk through the house with the help of sensors and clean the mess, following before running out the battery, to reach the dock to recharge.

In the future, the refrigerator will know what foods stores, when foods expire and how much should be purchased to complete the assortment preseted "stock" and why not, to order them.

The ovens will have memory database with cooking instructions and the user will scan a bar code that will be a recipe, following that the device to prepare itself one dish, based on instructions from memory.

The house windows will be made of liquid crystal so that it can be adjusted brightness inside without resorting to louvers, only by changing the crystal orientation.

Certainly there will be many inventions that will revolutionize the future house. All you have to do is use your imagination to view any combination of automatic functions.

A. Automated house benefits

A home equipped with the latest automated systems should simplify our lives and no way to complicate it. In most cases the purpose of improving technology is to protect people and ensure their comfort.

Therefore, the integration of intelligent equipment in a residential infrastructure just makes life simpler and saves us time and money, indirectly. An intelligent house is one that is equipped with multiple systems for managing electricity, house maintenance costs, their payment through electronic systems, monitoring various ways, etc. The automation makes our house a practical multifunctional complex and easy to use, integrating harmoniously and coherently entire house audio and video, communication, cooling systems, lighting, security, internet, etc. We feel good to think that a "brain" is keeping monitoring our house to a pleasant temperature, the TV automatically starts when we enter the living room, the stereo starts at a certain time, the irrigation of our plants in the garden is made only when founds that the soil moisture has reached a low value. All these elements are common in a "digital home". These goals can be achieved due to careful software developed by specialized companies.

These programs are made on the idea of connecting separate or sequential use of such equipments

Therefore they must be constantly adaptable, willing to support new and new requirements as they arise.

For this reason, does not exist the condition of a direct dependence of a particular hardware manufacturer

We need to not forget that we use in our house equipments and accessories produced by different companies, based on diverse technologies.

For this reason it is imperative that the installed software and hardware to be able to assimilate these technologies and then control them without any difficulty.

The "House Brain" can control the relays ordered various types of sensors (temperature, motion, lighting, smoke, etc..) surveillance cameras, audio-video equipments, power or calorie counters and more.

The common element of all these features is that it can be connected to a computer and controlled by it.

Monitoring of such equipments requires besides turning on lights, opening doors, triggering certain more complex operations and the management of certain crisis situations in the system, without the intervention of the owner.

The flow diagram of data collection and processing without connecting to a server is shown in Fig. 4.

The possible controlled functions of the intelligent buildings could be as following: communication, fire&safety, high speed internet, security, access, water, energy, light, elevator, IP/ digital TV, 24/7 Monitoring, IP Telephony, HVAC.

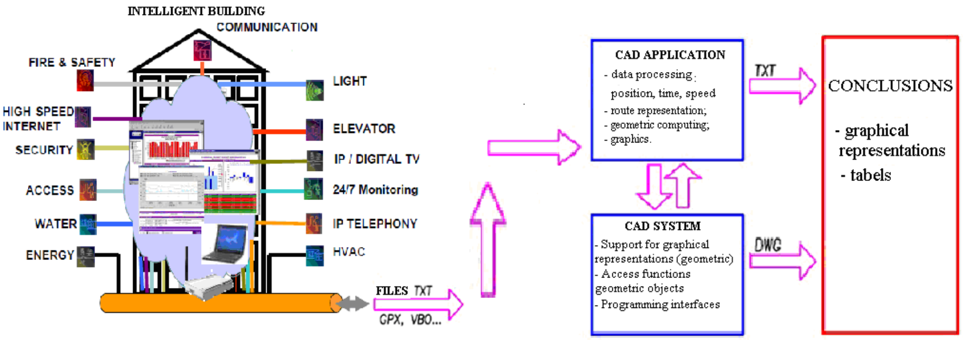


Fig.4. Flow diagram of data collection and processing, without connecting to a server

If we are connecting the intelligent building to a WebDAV server the system could trigge certain more complex operations and the management of certain crisis situations in the system, without the intervention of the owner Fig. 5.

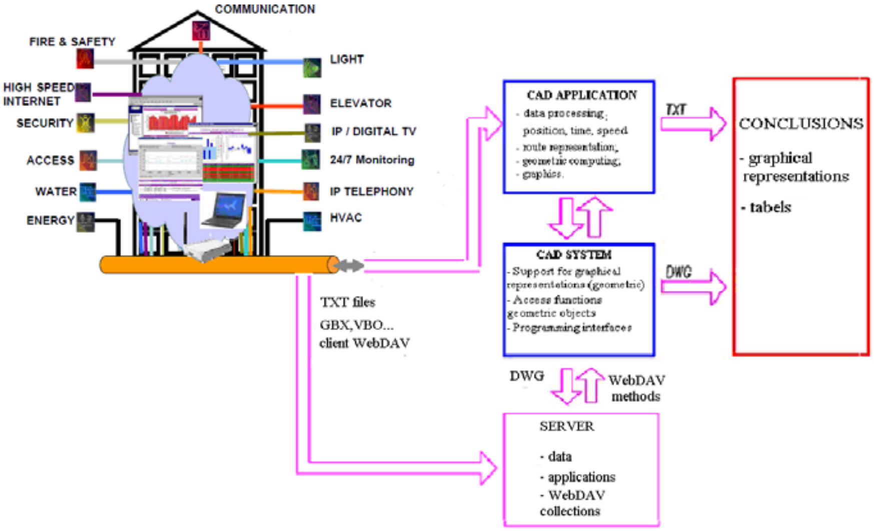


Fig.5. Flow diagram of data collection and processing, to connect to a WebDAV server

V. Study subject building

Our study is a study subject building that meets the highest standards of quality, from the point of view of functionality, image arts, architecture, materials and technologies used. The building harmonizes four essential components of sustainable construction, namely:

- Ecological component - the use of the conventional energy sources, reduced pollution, energy saving, reused resources (water, air), energy recovery, optimizing consumption, environmental protection.

- Economic component - the investment value, operational maintenance cost, operating efficiency, the use of advanced and sustainable materials, design innovation;

- Social component - ensuring optimal working conditions, health, hygiene, comfort in use, indoor environmental quality.

- Cultural component - the conditions for quality construction, matching in the territory, urban image, design innovation and new design solutions.

The building consists of ground floor, floor and semi-basement which will house offices, toilets, horizontal and vertical access ways, classrooms, a conference room on the ground floor, and at basement, a data center, civil defense shelter, workshops and technical staff.

The truly effective design process energy comes from the growing need to manage the whole building as an integrated system.

The design is a team effort, an effort that includes investor, architect and specialized engineers and, with a growing recognition, on those involved in project execution.

This project involved a considerable level of coordination and collaboration.

Also this building is one that meets the "green building" principle.

In acceptance of RoGBC (Romania Green Build Professional), a" green building" means a building which involves a high quality of workmanship, high energy efficiency and a low environmental impact.

A green or “sustainable” building “meets the needs of the present generation without compromising the ability of future generations to meet their own needs”.

That building incorporates strategies in its design, construction and operation in order to reduce energy use and minimize or eliminate negative impact on the planet [4].

VI. The project and Computer Aided Design (CAD) concept

The project is a "symbol marking the organization's transformation into a responsible company"

Project homogeneity design stages depend heavily on their incorporation into a whole. The built environment is a particular manifestation of technological innovation and the modes by which we apply technology in the design, construction and use has direct implications on energy consumption.

To achieve an efficient and quality product, Computer Aided Design (CAD) seems to be indispensable.

Working optimal method chosen for the first phase of building design is: 2D lines, 3D components work or a hybrid model construction. There are all standard interfaces for efficient data exchange with other programs (over 50 in number including DWG, DXF, DGN, IFC, PDF).

Any building element was modeled in detail - from the foundation, continuing with mono or multilayer components like walls, ground floor, floors and semi-basement, columns, beams, stairs and balanced standard, up to a variety of roof shapes. Interoperability without limits begins with a 2D design of any building level (Fig. 6).

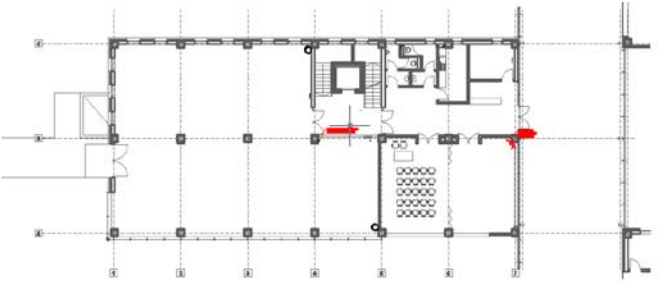


Fig.6. 2D architectural plan of the building level

In the above figure is the 2D architectural plan of the building level, achieved in AutoCAD Architecture 2009. In Fig. 7 are the longitudinal section plans and building cross section. The drawings are also achieved in AutoCAD Architecture 2009.

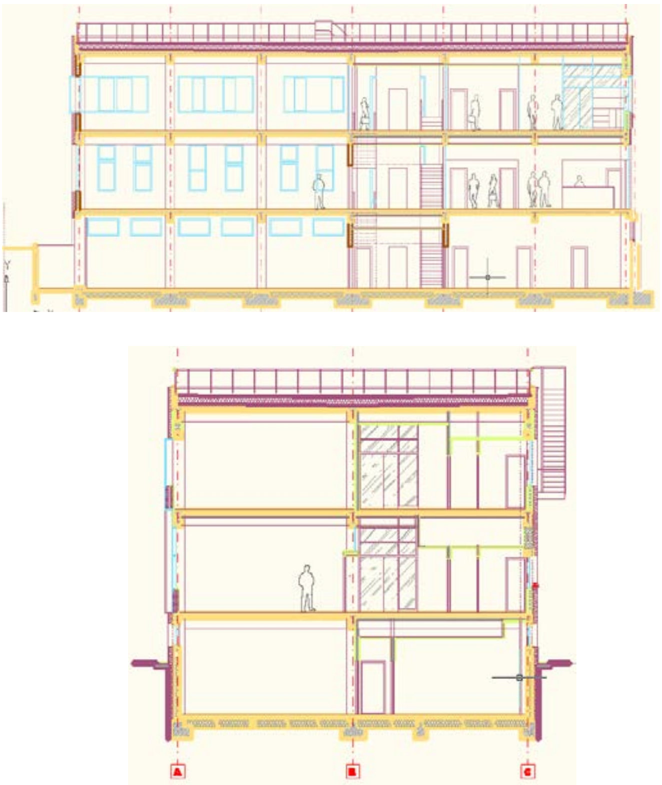


Fig.7. The longitudinal section plans and building cross section achieved in AutoCAD Architecture 2009

AutoCAD® Architecture software is the version of

AutoCAD specifically for architectural design. Architectural drafting tools enable you to design and document more efficiently in the familiar AutoCAD environment. Start working in AutoCAD Architecture and experience productivity gains right away, while learning new features at your own pace.

AutoCAD Architecture 2009 software suite from Autodesk has ingenious design and practical functions such as a pattern of windows and doors, architraves and Camcorder functions for advanced functions for facades (walls, curtain) or railings and fences. AutoCAD Architecture 2009 is a program that supports the entire design process from the first sketches of architecture to achieve overall design and execution drawings for reinforcement.The holistic design begins with the implementation of sustainable strategies and consider life-cycle analysis of the whole building.

This approach is essential to maximize individual potential and innovative technologies [5].

At data exchange level, using the explicit coding machine readability, the building model supports the automatic data transfer, improving the availability of design information for other users throughout the design and building processes which will happen later.

These new relationships between data, drawings and analyzes are used to reduce the number of errors, inconsistencies and ambiguities in the drawings produced by the structural engineer for transmission to the rest of the design team.

The piping 3D model of HVAC system is shown in Fig 8.

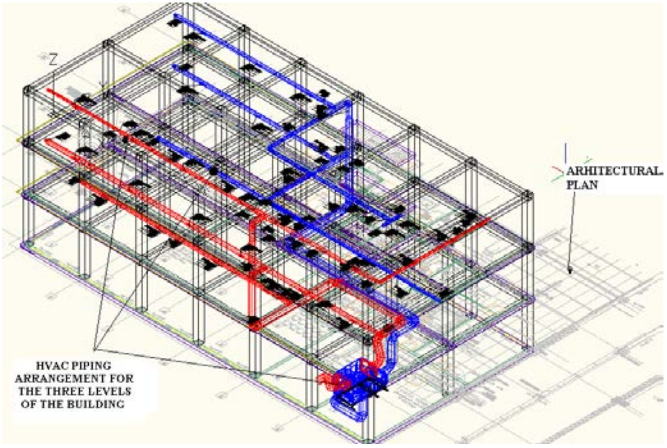

Z
ARHITECTURAL PLAN
HVAC PIPING ARRANGEMENT FOR THE THREE LEVELS OF THE BUILDING

Fig.8.The piping 3D model of HVAC system

Also, all HVAC piping systems have been set up in AutoCAD Architecture 2009.

The structural model plays a central role in this process. The effective coordination and review meetings occur when architectural; structural and installations models are available to be combined into a single virtual model. In Fig. 9 are represented the cable routing for IT equipment and electric cables in wireframe mode type (Fig. 9 a) and the cable routing for building level in shade mode type (Fig. 9 b).

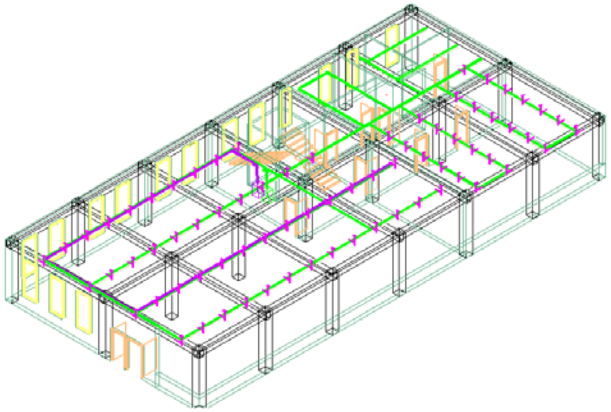


Fig.9a.The cable routing for IT equipment and electric cables in wireframe mode type

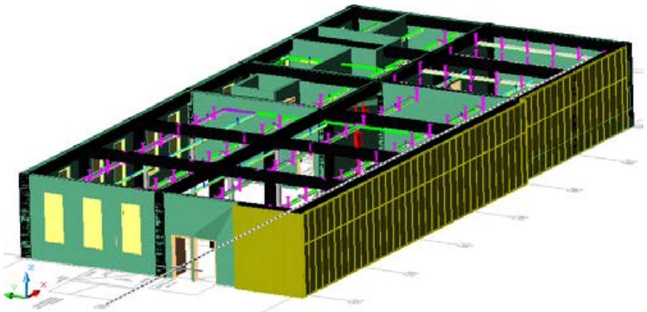


Fig.9b.The cable routing for building level in shade mode type

The cable routing was designed by CAD-2D dwg files of architectural plans of building levels.

For such buildings are important both facilities with all categories installations and the architect ability to think enough flexible space to allow its adaptation to any type of activity.

In this project, using MATLAB programming environment, was performed electrical power and lighting sizing and was calculated earthing resistance.

VII. MATLAB

MATLAB (MATrix LABoratory) is an interactive program for processing numerical data provided in a vectorial or matrix form.

MATLAB includes specific applications called TOOLBOX sites. These are extensive collections of MATLAB functions that are developing the programming environment from one variant to another, in order to solve specific problems. In the case of signal processing we will work mostly with "Signal Processing Toolbox".

MATLAB works with programs contained in the files. Files which are containing MATLAB instructions are called M file (its have extension .m). M files can be regarded as macros of MATLAB commands saved in the files with the extension *.m*., ie namefile.m.

An *m.file* can either be a function of input and output variables or a list of commands.

A MATLAB program can be written as script files or as a function files. A script file is an external file that contains a sequence of MATLAB commands

After full implementation of a script file, the created variables by this type of files remain in the application memory. If the first line of the file contains the word "function", that file is a function file, which is characterized by the fact that it can work with arguments.

At the end of the execution of a function, in computer memory remains only its output variables.

MATLAB works with two types of windows: a command window and a graphics window. At a time we can open only a command window.

The graphics window is used in graphic representation of the data. Multiple graphics windows may be opened in the same time

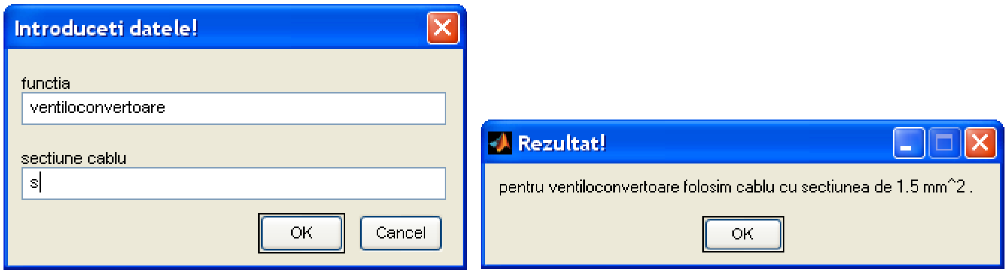
For electric power sizing were used as general data entry as following: the number of receivers (n) correction factor (cn), power factor (cosfic) demand coefficient (kc) section in mm2 (s), the intensity depending on the conductor section (I), network voltage (Un), number of air vents (x, y), the installed capacity (Pid), output power (Putild), power consumption (Pc), the intensity of current consumption (Ic). Using formulas and comparisons we obtain as output the selected conductor cross section, in mm2.

The floor electrical installation force consists of: air vents, central air handling units, power air conditioning, power rack, central fire detection, forced entry central, circuit sockets and backup circuits (Table 1).

Table 1

air vents	
x1=15;	% number of air vents
v1=7*50; %[W]	
v2=8*90; %[W]	
Pid=v1+v2;	% installed capacity
Putilvcv=1600;	% output power
cnvcv=interp1(n,cn,x1);	% Correction coefficient;
kcvcv=kc(2)+((1-kc(2))/cnvcv);	% corrected demand coefficient;
Pcvcv=kcvcv*Piv;	% power consumption;
Icvcv=Putilvcv/(Un*cosfic(2))	% intensity of current consumption
if Icvcv < ICu(1)	
svcv=s(1)	% conductor cross section [mm2]
end	

In the Fig. 10, below is shown a sequence of Matlab software for sizing floor electrical power.



functia

ventiloconvertoare

sectiune cablu

s

pentru ventiloconvertoare folosim cablu cu sectiunea de 1.5 mm^2 .

Fig. 10. A sequence of Matlab software for sizing floor electrical power

Depending on the current running through the wire, using Table 2, we choose appropriate circuit breaker.

Table2. The currents standardized values for automatic switches

<table>
<tr><td>Rated current [A]</td><td colspan="4">6, 10, 16, 25, 32, 40, 50, 63, 80, 100, 160, 250, 400, 630, 800, 1000, 1250, 1600, 2500.</td></tr>
<tr><td>Nominal voltage [V]</td><td colspan="2">230</td><td colspan="2">400</td></tr>
<tr><td>Numarul de poli</td><td>Monopolar 1P+N</td><td colspan="2">Bipolar 2P+N</td><td>3P+N</td></tr>
</table>

According to Table 2 we have chosen a circuit breaker of 10 A.

VIII.INTEGRATION OF LIGHTING SYSTEMS INTO COMPLEX TECHNICAL INTELLIGENT BUILDING

Our building is seen as a set of systems and a special attention was paid to the lighting system

Any construction must meet a number of requirements summarized in three key factors that are working each other: conception, design and their composition.

Thus we have three factors:
- Human factor - which requires achieving the necessary comfort conditions for human activity, such as: temperature, humidity, lighting, noise, etc.. These depend on the type of work the people perform during building construction activities.
- Human activity factor - requires construction functional composition to meet the requirements of the activity type. This design construction of a residential building differs for automobile production or for buildings constructios.
- Nature factor - involves all actions resulting from interaction design - environment relating to: the seismicity level of the area, the intensity of climate actions (wind, snow, rain, frost, etc.), the quality of the soil foundation, groundwater levels, groundwater aggressiveness etc..

The stated these factors accounted for over time into "construction laws" as: interim technical guidance, technical guidelines, design manuals, design standards, technical regulations, etc.

Lighting optimization on the building levels was performed with the software DIALux DIALux lighting-design tool and the utility software ABS Autodesk Autodesk Building Systems 2006-UK - AdeskUKABSUtils.arx & AdeskUKABSUtilsEnu.dll has established the arrangement of lightings and their number.

The positioning of light sources on the building floor level the using the 3D model is shown in Fig. 11.

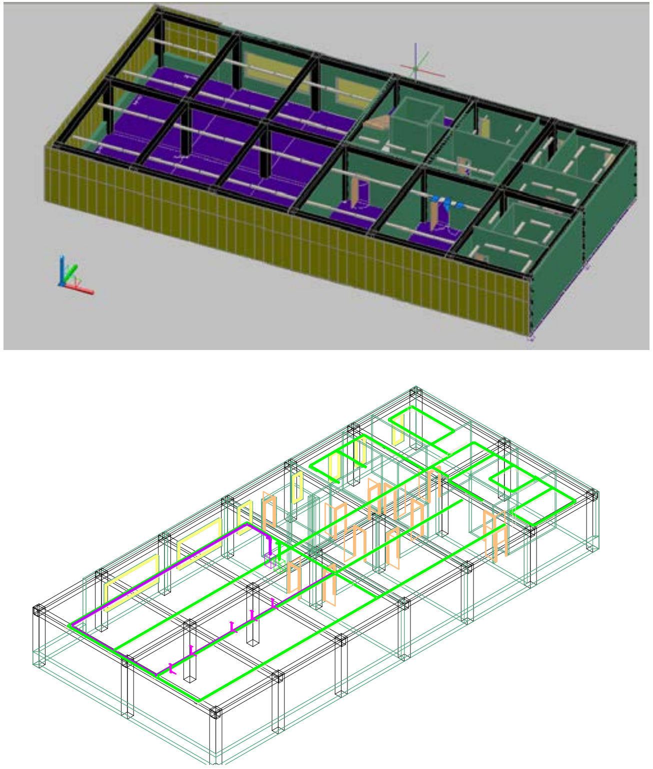


Fig.11. Positioning of light sources on the building floor level the using the 3D model

The positioning of light sources on the building floor level the using the 2D model is shown in Fig. 12.

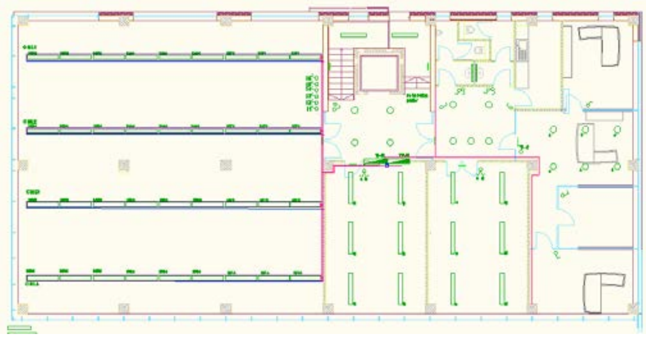


Fig.12. Positioning of light sources on the building floor level the using the 2D model

In order to create a energy efficient building, we need to imply intelligent design, right from the beginning of the project.

IX. Conclusion

The global approach in terms of the intelligent design of buildings concept is clearly subordinated to economic criteria expressed more or less in a rigorous form of terms such as: increasing labor productivity of those who are working in such buildings, administrative improvement, cost reduction on information technology and communications, insurance of develped activities, obtaining increased intercooperation facilities, improving ergonomic requirements and other requirements relative to human activities (culture, entertainment, sports)

Within the intelligent buildings, the intelligent management and control solutions include convergence of building management and energy with security, life safety and fire safety, communication, IT equipment rooms for IT equipments, machines automation and vertical-specific applications to create a highly adaptable, durable and economical building.

Graphics are engineers creative works and are, traditionally, results of translated based on CAD drawings. Currently, the collaborative design makes a more sophisticated sense, when referring to the design of the built environment. Organizations can improve their efficiency and productivity through collaboration contacts and collective intelligence networking and continuous virtual communication between team members.

References

[1] Dr. Jane L. Snowdon J. Jones. (2009, June 10). IBM Smarter Energy Management Systems for Intelligent Buildings. Available: http://citris-uc.org/files/Snowden%20IBM%20Research%20061009.pdf

[2] Elena Rastei, Romania Green Building Council (2012, March 7).Cladiri Verzi: Proiectare Sustenabila. Design Integrat. Available:http://www.ecomagazin.ro/cladiri-verzi-proiectare-sustenabila-design-integrat/

[3] Cristian CIUREA, “A Metrics Approach for Collaborative Systems,” Informatica Economică, vol. 13, no. 2, pp. 41–49, Feb. 2009.

[4] Paul Rinder, RoGBC Green buildings.Principles, assessment and certification, design. Available:http://www.euroconferinte.ro/prezentari/Tema107.pdf

[5] http://www.autodesk.com/products/autodesk-autocad-architecture/overview.

Integrated Development Environment for Remote Application Platform

Eclipse Rap – A Case study

Sagaya Aurelia1, Xavier Patrick Kishore, Omer Saleh

Abstract-An integrated development environment (IDE) (also known as integrated design environment, integrated debugging environment or interactive development environment) is a software application that provides comprehensive facilities to computer programmers for software development. Eclipse is a community for individuals and organizations who wish to collaborate on open source software.

Eclipse Remote Application Platform (RAP 2.1.0M2) is a framework for modular business applications that can be accessed from different types of clients including web browsers, rich clients, and mobile devices. This paper reviews and analysis Eclipse RAP and its features.

Keywords-- IDE; Eclipse; RAP; RWT

I. Introduction

ANintegrated development environment (IDE) is a programming environment that has been packaged as an application program, typically consisting of a code editor, a compiler, a debugger, and a graphical user interface builder. The IDE may be a standalone application or may be included as part of one or more existing and compatible applications. Eclipse projects are focused on building an open development platform comprised of extensible frameworks, tools and runtimes for building, deploying and managing software across the lifecycle. In general, the Eclipse provides four services 1) IT Infrastructure, 2) IP Management,3) Development Process, and 4) Ecosystem Development. Eclipse Remote Application Platform (RAP) provides a powerful, multi-platform widget toolkit with SWT API that enables developers to write applications entirely in Java and re-use the same code on different platforms.

The paper proceeds as follows in section 2, we will present about IDE. In Section 3, we will present Eclipse IDE. We follow in Section 4 with Eclipse RAP its architecture, life cycle phase, protocols how it works as a server and for embedded system. Finally suggestions along with conclusions are stated in section 5.

II. Integrated Development Environment

IDE isan integrated development environment, the handy, dandy piece of software that acts as text editor, debugger and compiler all in one sometimes-bloated but generally useful package [11]. Most common features, such as debugging, version control and data structure browsing, help a developer quickly execute actions without switching to other applications. Thus, IDE helps maximize productivity by providing similar user interfaces (UI) for related components and reduces the time taken to learn the language. An IDE supports single or multiple languages [12].

Selecting a good IDE is based on factors, such as language support, operating system (OS) needs and costs associated with using the IDE etc. Visual Studio, Delphi, JBuilder, FrontPage and DreamWeaver are all examples of IDEs. There are so many features an IDE can contain that the following list contains only a selected few[9].

A. *Code completion or code insight:* The ability of an IDE to know a language's keywords and function names is crucial. The IDE may use this knowledge to do such things as highlight typographic errors, suggest a list of available functions based on the appropriate situation, or offer a function's definition from ` official documentation.

B. *Resource management:*When creating applications, languages often rely on certain resources, like library or header files, to be at specific locations. IDEs should be able

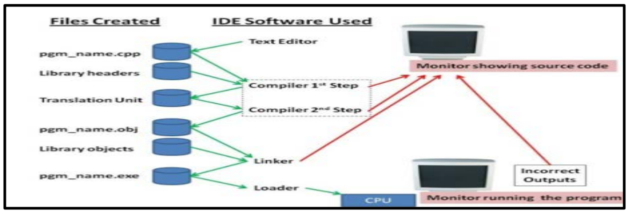


Fig. 1 The role of IDE in development stage [19]

.

C. *To manage these resources.* An IDE should be aware of any required resources so that errors can be spotted at the development stage and not later, in the compile or build stage.

F. A. Dr. Omer Saleh, Department of Computer Science, Faculty of Education, Beniwalid, Libya (immer.jomah@gmail.com)

S. B. Xavier Patrick Kishore, Department of Computer Science, Faculty of Education, Beniwalid, Libya (Patrick.kishore@gmail.com)

T. C. P. Sagaya Aurelia1, Department of Computer Science, Faculty of Education, Beniwalid, Libya,(psagaya.aurelia@gmail.com)

D. *Access Databases:* To help connect Java applications to databases IDEs can access different databases and query data contained within them.

E. *Optimization:* As Java applications become more complex, speed and efficiency become more important. Profilers built into the IDE can highlight areas where the Java code could be improved.

F. *Project management:* This can be twofold. First, many IDEs have documentation tools that either automate the entry of developer comments, or may actually force developers to write comments in different areas. Second, simply by having a visual presentation of resources, it should be a lot easier to know how an application is laid out as opposed to traversing the file system for arcane files in the file system.

III. ECLIPSE –IDE

Eclipse is a universal platform for integrating development tools.Eclipse is the free and open-source editor upon which many development frameworks are based. The overview of Eclipse is shown in figure 2. Eclipse began as a Java development environment and has greatly expanded through a system of lightweight plugins. Eclipse is created by an Open Source community and is used in several different areas, e.g. as a development environment for Java or Android applications. The Eclipse Open Source community has over 200 Open Source projects covering different aspects of software development [10].

The Eclipse ++ can be extended with additional software components. Eclipse calls this software components *plug-ins*. Several Open Source projects and companies have extended the Eclipse IDE[18]. The extended overview of eclipse is shown in figure 2.

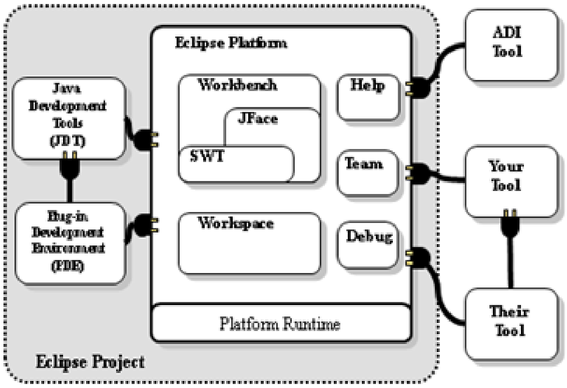


Fig. 2. Overview of eclipse [10]

IV. ECLIPSE BASED APPLICATIONS [20]

An Eclipse application consists of individual software components as shown in figure 3 and 4. The Eclipse IDE can be viewed as a special Eclipse application with the focus on supporting software development.

The components of the Eclipse IDE are primarily the following. Please note that the graph should display the concept, the displayed relationship is not 100 % accurate.

OSGi is a specification which describes a modular approach for Java application. Equinox is one implementation of OSGi and is used by the Eclipse platform. The Equinox runtime provides the necessary framework to run a modular Eclipse application.

SWT is the standard user interface component library used by Eclipse. JFace provides some convenient APIs on top of SWT. The workbench provides the framework for the application. The workbench is responsible for displaying all other UI components.

On top of these base components, the Eclipse IDE adds components which are important for an IDE application, for example the Java Development Tools (JDT) or version control support (EGit).

On top of these base components, the Eclipse IDE adds components which are important for an IDE application, for example the Java Development Tools (JDT) or version control support (EGit).

Eclipse 4 has a different programming model then Eclipse 3.x. Eclipse 4 provides the *3.x Compatibility Layer* component which maps the 3.x API to the 4.0 API. This allows Eclipse 3.x based components to run unmodified on Eclipse 4.

Eclipse based applications which are not primarily used as software development tools are called Eclipse RCP applications. An Eclipse 4 RCP application typically uses the base components of the Eclipse platform and adds additional application specific components.

The programming model of OSGi (Equinox) allows you to define dynamic software components, i.e. OSGi services, which can also be part of an Eclipse based application.

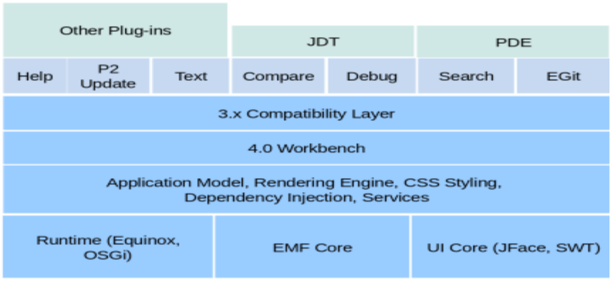


Fig. 3 Eclipse based application [20]

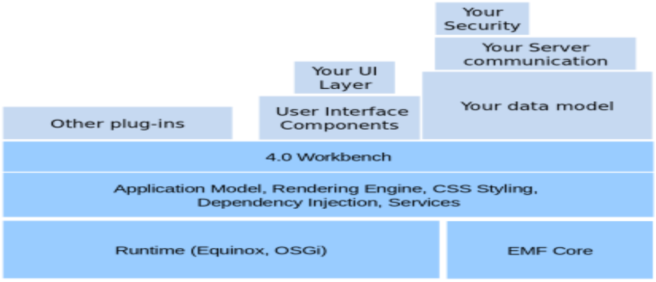

Your Security
Your Server communication
Your UI Layer
Your data model
Other plug-ins
User Interface Components
4.0 Workbench
Application Model, Rendering Engine, CSS Styling, Dependency Injection, Services
Runtime (Equinox, OSGi)
EMF Core

Fig. 4. Eclipse based application [20]

V. The Eclipse Platform Provides a Tool Integration Framework [8]

The Eclipse Platform reduces the cost of tool integration by providing a large number of services, APIs, and frameworks that enable effective and scalable tool integration. Wherever possible, Eclipse uses open standards to limit tool vendor investment and reduce time to market. The Platform provides a focal point for integrating and configuring best-of-breed tools in a manner that best fits the end user's development process and Web application architecture. The Eclipse Workbench provides a central integration point for project control and an integration mechanism for resource-specific tools. The Eclipse Platform can also provide services common to different tools including user interface frameworks, managing relationships between components, component version management, and publishing services. Using Eclipse simplifies tool integration by allowing tools to integrate with the platform instead of each other. This significantly reduces the number of integration problems that must be solved, and provides a more consistent environment for the end user[8].

The eclipse IDE consists primarily of plug-ins built on the Eclipse base. A typical development system has more than shown here. Building a non –IDE application on Eclipse is a matter of adding

Application specific plug-ins as shown in figure 5. Developers ultimately gain a plug-in architecture and a common graphical interface.

VI. Eclipse RAP Remote Application Platform

The Remote Application Platform (RAP) formerly Rich Ajax Platform is a framework for modular business applications that can be accessed from different types of clients including web browsers, rich clients, and mobile devices as shown in figure 6. It provides a powerful, multi-platform widget toolkit with SWT API that enables developers to write applications entirely in Java and re-use the same code on different platforms. It enables developers to build rich user interfaces using the Eclipse tools and common APIs [4].

Regardless of the client platform, RAP applications run on a server that communicates with its clients over HTTP.

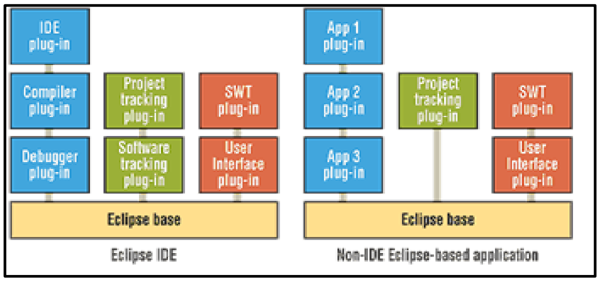
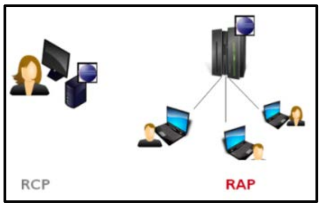


Fig. 5 Plugin of Eclipse IDE and Non IDE Eclipse based application. [1]

It can be considered a counterpart for web development to the Rich Client Platform (RCP). RAP encourages sharing source code between RCP and RAP applications as shown in figure 7 to reduce the development effort for business applications that need both desktop-based and web-based front ends.

VII. RAP Architecture

RAP is to the web as RCP to the desktop. It inherits all the goodness from RCP such as workbench extension points model, event-driven SWT/JFace APIs, and componentized OSGi design. As indicated in figure 8 (a), the only difference between the architecture of RAP and that of RCP is the implementation of SWT/RWT. RWT is actually a bundle providing web-specific implementation of SWT's widgets based on the qooxdoo toolkit. In RAP, almost no SWT API is changed [15].



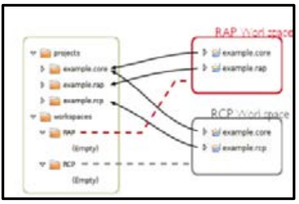


Fig. 6. RAP is a MultiuserFig. 7. Sharing Projects

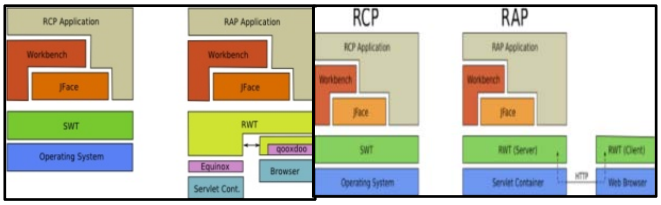


Fig.8 (a) RAP ArchitectureFig. 8 (b) Architecture of RAP in Server Centric Framework

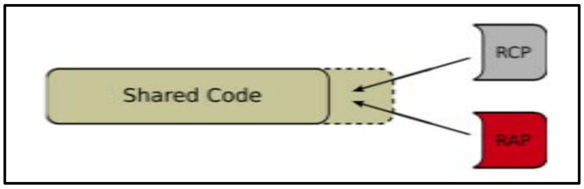


Fig. 9 a). Fragments

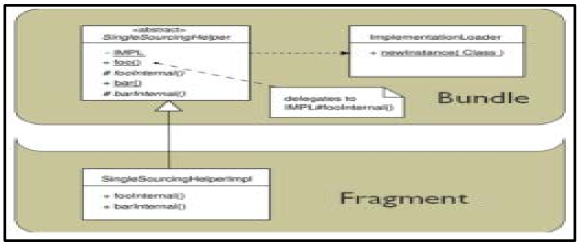


Fig. 9 b). Delegation

As indicated in figure 8(b), in a server-centric framework, the application is hosted on the application server and all processing is done on the server. The client browser is only used for data presentation. Consequently, it leaves small footprints on browsers: it waits for instructions from the server to create corresponding widgets on demand [15].

When the data has to be shared fragments are added to the shared codes. Fragments are a sort of patches as shown in figure 9a. The concepts of fragments are used both in RCP and RAP. Some of the fragment properties are fragment id, version, name, provider and the class path and the host plugin also has properties as plug-in ID, minimum and maximum version. Fig. 9b shows the delegation process where after the fragment process it is bundled together.

Like building a RCP application, building a RAP application is a process of building plug-ins and bundles: On the UI side, contributing widget plug-ins; On the server side, since it is powered by server-side Equinox, contributing servlet plug-ins. Unsurprisingly, it also inherits the benefits of any OSGi application such as dynamically adding/removing bundles[15]. As shown in the below figure eclipse rap works as server figure 10 and also for embedded system as shown in figure 11. It provides a complete target platform based onEquinox, including subsets of SWT, JFace, and Workbench APIs. With the RAP OSGi integration, they can be composed ofmodules and using the OSGi service model. The core library can also be used in traditional web applications without OSGi [16].

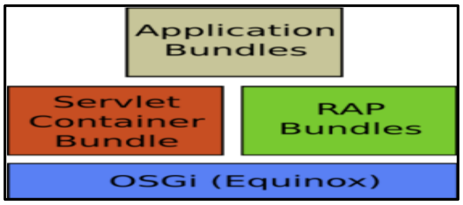


Fig 10. Eclipse RAP as server

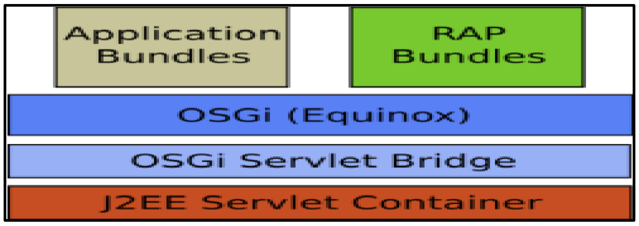


Fig. 11 Eclipse RAP for embedded system

Applications or components (coming in the form of bundles for deployment) can be remotely installed, started, stopped, updated, and uninstalled without requiring a reboot; management of Java packages/classes is specified in great detail. Application life cycle management (start, stop, install, etc.) is done via APIs that allow for remote downloading[14].

Thedefaultwebclientuses JavaScripttorendertheUIinthe browser.

VIII. Eclipse RAP consist of [1]

1. *Widget Toolkit*

With RAP, you don't create UIs with HTML and browser technologies, but with the Java API of SWT, the widget toolkitused in Eclipse. The RAP Widget Toolkit (RWT) provides a comprehensive set of powerful SWT widgets, also including layout managers and event listeners. RWT architecture is shown in figure 12.

1.1 The RWT Lifecycle Phases[5]

The phases are: as follows and shown in figure 13

a) *Prepare UI Root*: Responsible for invoking entry points.

b) *Read Data* :Reading request parameters and applying the contained status information to the corresponding widgets. As an example, if a user has entered some characters into a Text control, the characters are transmitted and applied to the text attribute of the Text instance.

c) *Process Action* : Events are processed which trigger user actions. As an example, when a Button has been pushed, the Selection Listeners attached to the Button are called.

Render: JavaScript code is generated for the response that applies the state changes to the client. Only those

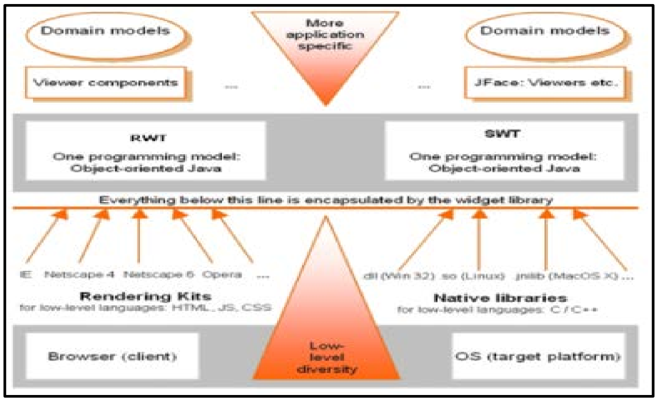

Domain models
More application specific
Domain models
Viewer components
...
...
JFace: Viewers etc.
RWT
One programming model: Object-oriented Java
SWT
One programming model: Object-oriented Java
Everything below this line is encapsulated by the widget library
IE Netscape 4 Netscape 6 Opera ...
Rendering Kits
for low-level languages: HTML, JS, CSS
dll (Win 32) so (Linux) jnilib (MacOS X) ...
Native libraries
for low-level languages: C / C++
Browser (client)
Low-level diversity
OS (target platform)

Fig . 12 RWT Architecture[17]

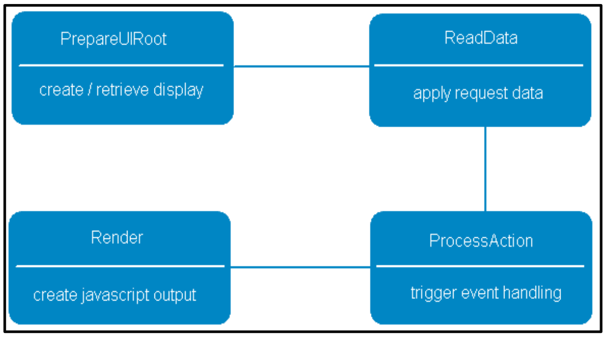


Fig. 13. RWT Life Cycle Phases

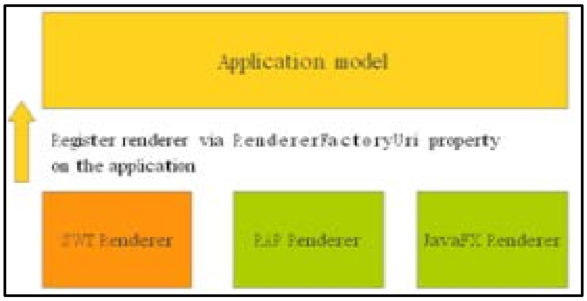

Application model
Register renderer via RendererFactoryUri property on the application
SWT Renderer
RAP Renderer
JavaFX Renderer

Fig. 14Creating your own renderer [3]

Fig. 15Renderers for different platform

widget attributes that were changed during the processing of the current request are being rendered. This results in a minimal amount of data that needs to be transferred to the client.The widget tree is not manipulated in this phase anymore.

User can create their own desired as shown in figure 14 and sample renderers for different platforms are shown in figure 15.

2. Cross Platform[2]

The default RAP Web client supports these browsers:

- Internet Explorer 7+
- Google Chrome 7+
- Firefox 3.5+
- Safari 4+
- Opera 10+
- iOs 5+
- Android 3 (Limited)

No browser plug-ins is required by the default client, only JavaScript needs to be enabled. However, custom widgets are free to build on any third-party API. Support on mobile browsers has some limitations. Other platforms can be supported by alternative RAP clients connecting to RAP's open protocol.

2.1RAP Protocol [2]

The term RAP Protocol is used to describe the JSON-based message format used in the HTTP communication between a RAP client and a RAP server. This protocol was introduced in RAP 1.5 to replace the plain JavaScript responses that have been used in previous releases. As of RAP 2.0, the entire communication between client and server uses this protocol. The idea of the RAP Protocol is to fully decouple RAP server and RAP client with the intention of making the client exchangeable. At this point of development, the most interesting use cases have been RAP clients for mobile devices (Android and iOS), written in the client's native language. As a consequence, all logic that is specific to the default RAP client has been moved from the server to the client.

The protocol does not only enable clients in other programming languages, it also opens the door to a new class of applications – applications that need to address a wide range of hardware from desktops to specialized devices (e.g. mobile data entry or point of sales solutions). Or applications that require integration with attached hardware devices. We think that this is a major new achievement for RAP warranting a major release – and a feature that sets RAP apart from other frameworks[6].

3. Integration

Making it possible to integrate RAP with other Java technologies is one of our main objectives. We're doing so by making RAP compatible with JEE and OSGi and by limiting dependencies to the necessary minimum. A partial list of compatible technologies:

- RAP applications can be deployed directly as OSGi bundles
- The JEE compatibility mode in RAP makes it possible to use clustering
- Equinox Security Integration ensures your data is safe at all times.

4. SingleSourcing

RAP allows toaddress different platforms with a shared code base. Applications can be developed for the desktop, the web browser, and even mobile clients without duplicating code. Due to RAP's high compatibility with the Eclipse UI technologies, a lot of existing code targeted at the desktop can be re-used for the user’s web application with minimal changes.

Typical scenarios are:

- Porting an existing SWT/RCP application to the web.
- Developing a new application that can run on the desktop and in the browser.
- Re-using existing libraries developed for previous applications.
- Utilize RCP-compatible open source libraries and frameworks. Open Source

Notable Additional Features[2]

- *Client Class and Client Services*

All features specific to the RAP client (which is exchangeable as of RAP 2.0) are handled by the client class and the client services. This includes support for browser history, JavaScript execution and retrieving the clients time zone offset.

- *HTTP File Upload*

Unlike SWT, RWT cannot simply access the user's file system and read data from it. As an alternative, the File Upload widget can be used. The widget looks like a button, but when clicked will open the file picker dialog of the user's browser. After a file has been selected, it can programmatically be send to any HTTP server.

- *Fixed Columns*

It is possible in RWT to exclude some columns from Tree or Table from scrolling.

- *Multi-User Environment*

RAP operates in a multi-user environment and provides some additional API that helps dealing with the consequences. Notable Limitations

Notable Limitations[2]

- Few Unimplemented Features
- Unimplemented Widgets such as StyledText, Tracker, TaskBar, Tray
- Painting Limitations Some methods are unimplemented, including copyArea, drawPath, setClipping, setTransform, setInterpolation, setLineDash and setXORMode

- Limitations in Dialogs: Dialog, ColorDialog, FontDialog, MessageBox
- Limitations of the Browser widget :Since the Browser widget is based on the HTML iframe element, there are some restrictions
- Limitations in Mouse and Key Events
- Limitations in Verify and Modify Events
- Limitations in Drag and Drop
- Limitations when using background threads

IX. Conclusion

But integrated tools require a platform of services, frameworks, and standards that allow vendors to focus on their value-add while reusing common infrastructure. The platform must include a workbench that provides a common view of the whole application across all resource types and the entire team. And the platform must be accessible to tool vendors under an acceptable license. Eclipse not only provides such a platform, but its architecture also provides flexibility in how tool venders integrate their tools and at what level. This allows vendors to match their integration investment with their product needs and market window.

For simple integration, use invocation integration to provide users navigation, access, editing, and management of file-based resources. Use data integration to share data between tools that are otherwise unconnected [8]. When data integration isn't enough, use API integration to provide secure access to encapsulated data. Eclipse RAP

Is not only useful for software industry but also for academic. It can be used to integrate all colleges and furthermore all universities to introduce a centralized education system.

References

[1] www.electronicdesign.com
[2] www.eclipse.org/rap
[3] http://www.eclipsecon.org/europe2012/sites/eclipsecon.org.europe2012/files/Eclipse4_RAP.pdf
[4] Http://developer.eclipsesource.com/technology/crossplatform/#rap
[5] http://download.eclipse.org/rt/rap/doc/2.0/guide/reference/api/org/eclipse/rap/rwt/lifecycle/ILifeCycle.html
[6] http://eclipsesource.com/blogs/2012/11/26/rap-becomes-the-remote-application-platform/
[7] http:R//eclipse.org/rap/developers-Guide/devguide.php?topic=scopes.html&version=2.0
[8] http://www.eclipse.org/articles/Article-Levels-Of-Integration/levels-of-integration.html
[9] http://salfarisi25.wordpress.com/2010/12/22/advantage-and-disadvantage-of-using-ide/
[10] www.eclipse.org/
[11] http://mashable.com/2010/10/06/ide-guide/
[12] http://en.wikipedia.org/wiki/Integrated_development_environment
[13] http://help.eclipse.org/juno/
[14] http://en.wikipedia.org/wiki/OSGi
[15] http://owenou.com/2010/07/08/introducing-eclipse-rap.html
[16] http://dev.eclipse.org/mhonarc/lists/rt-pmc/pdfmFx7CiT7Ma.pdf
[17] http://www.pjug.org/docs/RAP.pdf
[18] http://en.wikipedia.org/wiki/Eclipse
[19] http://cnx.org/content/m18920/latest/graphics1.jpg
[20] http://www.vogella.com/articles/EclipseRCP/article.html

Dr. Omer Saleh MahmodJamah (January 25,1973) is now the Director of Post graduate cum Research and Development and Head of the department of Computer science, Faculty of education, Azzaytuna university, Baniwalid, Libya. He received his B.Sc. in Control System and Measurement (1995), M.Sc. in Electrical and Computer Measurement (2004), and Ph.D. in Electrical engineering, Automatics computer science and electronics from AGH University of technology, Krakow, Poland. He has done his Diploma in Planning and time management from Canada Global Centre, Canada. Now he is heading Computer Science department, Faculty of Education, Azzaytuna University, Baniwalid, Libya. His research interest includes multicriteria optimization for solving optimal control problems and Fuzzy logic. He has published 12 papers and attended various national and international Level conferences and workshops.

Mr. Xavier Patrick Kishore (November 6, 1973) received his BSc Mathematics (1994), Master of Computer Application (2002) and Diplomas in E-Commerce and Advanced software Technology. He has received Brain bench certification in Java and HTML. Now he is working in Department of computer science Faculty of Education, Azzaytuna University, Baniwalid, Libya. He is specialized in programming languages. His current research interest includes Natural language processing. He has authored more than 9 papers and attended many conferences.

Er. Mrs. Sagaya Aurelia(November 9,1978) par-time research scholar in Bharathidasan university . Now she is with department of Computer Science, Faculty of Education, Azzaytuna University, Bani-walid, Libya. She received her Diploma in Electronics and Communication (1997),B.E (Bachelor of Engineering specialized in Electronics and Communication Engineering(2000) and M.Tech in Information Technology(2004),she has also done her Post graduation diplomas in Business Administration (PGDBA) and Journalism and Mass Communication(PGDJMC). She has received Brainbench certification in HTML. Her current research interest includes Virtual reality, Augmented reality and Human Computer Interaction and User interface Design. She has authored14 papers and attendance several national and international level workshops and conferences.

Fusion of Visual and Acoustic for Active Acoustic Source Detection With Spatially Global GMM

R.Azzam and N. Aouf

Abstract— In this paper, we investigate the problem of reliable detection and localization of active sound source detection using a new fusion approach of the vision and the acoustic data for detection and localization. The usefulness of the solution is fundamental for both video surveillance and video conference systems. In this aim, we propose combining the two heterogeneous modalities of data by augmenting the 3-D vector of RGB colors used by the Spatially Global Gaussians Mixture Model (SGGMM) for background modeling and segmentation using the acoustic Data. The proposed model provides accurate detection of the interested targets and evaluation results using an implementation version on wireless sensors network (WSN) of the fusion approach shows performance improvement of the proposed detection and localization solution. This technique enabled a better detection of the moving acoustic source in comparison with the SGGMM only.

I. INTRODUCTION

Fast growing technology of WSN enabled the appearance of sensing devices with advanced features at low prices. At this asset researches in robust surveillance application propelled. A common approach for object detection and localisation is to imitate the cooperative functioning of human senses by combining both vision and hearing capabilities for better detection and position estimation of moving objects. Within this context, different solutions have been suggested.

In sensor networks systems, the fusion of vision and acoustic data is widely investigated: [1] has suggested a centralised architecture based on Extended Kalman Filter (EKF) to estimate the position of moving acoustic source. Throughout experimental results, the combination of the two types of data showed improvement in terms of target localisation accuracy. In [2], both acoustic and visual models are estimated as part of a joint unsupervised optimisation for speaker localization system. This method works through two steps: the first is for determining the number of speakers, while the second involve the usage of visual models to infer the location of the speaker in the video. In [3], author suggested the detection and localization of the active speaker by the fusion of visual reconstruction with a stereoscopic camera pair and sound-source localization using several microphones. Both devices are embedded into the head of a humanoid robot that works through statistical fusion model. This model associates 3D faces of potential speakers with 2D sound directions. In [4], a sensor fusion framework based on particle filters is presented. It proposes combining the detection and tracking results from a co-located acoustic array and video camera. The Particle filter based trackers are used to recursively estimate state probability density functions for the combined tracker. Overall target tracking performance is improved as the video controls the particles diversity at low signal-to-noise (SNR) levels of the acoustics. A video system was developed in [5] for ships identification and localisation based on the fusion of acoustic noise with video data. The fusion enabled the estimation of sound attenuation in a wide frequency band and the collection of a noise library of various ships. The latter is used for ship classification by passive acoustic methods.

For vision detection system, the background subtraction based techniques interest in detecting variation within scene across several image frames. The approach is based on comparing the current image with a reference one(s) of the background. Pixels of sharp variations are consequently classified into the foreground. Many techniques have been put to work using this principle among which we can list [6]: running average [7], Mixture of Gaussians [8], Mean shift [9], in addition to the Sequential KD Approximation [10]. The popularity of these techniques comes from their computational efficiency. Among the listed methods, the mixture of Gaussians (or the Gaussian mixture models) gained much respect due to its lower requirement of computation cost and reduced memory size allowed many applications to run in real-time at an acceptable performance in devices of lower processing budget such as the wireless camera sensors network. Two different approaches are possible with the segmentation using the GMM for foreground substation. The first considers modelling the historic values of each pixel on a mixture of Gaussians [8] using the Expectation-maximization algorithm (EM) for model parameter estimation. This modelling enables the detection of any sharp change in pixel value that exceeds a predefined probability threshold. The latter is to determine whether a pixel belongings to the background or a moving object. A second approach, [11] suggests clustering the pixels of the same frame in sets of closed values. The clusters are sorted in order of likelihood that models the background and adapted to deal with background and lighting variations. Incoming pixels are matched against their corresponding cluster group and are classified according to the matched cluster is considered part of the background.

Techniques of acoustic source localization vary according to situations. In the wireless sensors network two main approaches are used [12]: The first, is energy based source localization that is motivated by observation sound level decreases as the distance between the sound source and the listener becomes larger.The relation between the sound level and the distance from its source enable the estimation of the source location using multiple energy readings at different known source location. The second approach is a time difference of arrival (TDOA) based. It is motivated by the observation that the sound wave propagates at constant speed (sound speed) from the acoustic source to listeners. A number of microphones at known positions receive the propagated acoustic source at different times. Modelling of TDOA enables the estimation of the sound source position using different techniques from which EM [13], least squares (LS)[14] and weighted least square (WLS)[15] are among most used for acoustic localisation in WSN due to their low computational burden.

In this work, we propose a novel approach for active acoustic source detection by augmenting the 3D RGB SGGMM vector with the acoustic information. The proposed solution is implemented in a WSN for performance evaluation. The paper is organized as follows: Section 2 presents the SGGMM model, section 3 illustrates the acoustic localisation using the LS technique, section 4 presents how the SGGMM 3D vector is augmented with the acoustic signal. Section 4 describes the experimental setup while section 5 reports the results. The paper ends with a conclusion that summarizes the overall finding of the work.

III. The SGGMM Model

The SGMM works under the assumption that the camera is fixed. Each image pixel value is represented in feature space by a 3-D vector $x = [I_R, I_G, I_B]^T$. Additionally, the scene background is represented by a spatially global Gaussian mixture model of N Gaussians in 3-dimensional RGB colour space as follows:

$$p(x) = \sum_{i=1}^{N} w_i g_i(x_t, \mu_i, \Sigma_i) \quad (1)$$

where μ_i and Σ_i are respectively the spatial mean vector and covariance matrix of the *i-th* distribution. And w_i is an estimate of the weight, which reflects the likelihood that the corresponding distribution accounts for image colour and satisfies the criterion $\sum_{i=1,N} w_i = 1$.

Each Gaussian distribution $g_i(x_i, \mu_i, \Sigma_i)$ of the mixture is defined as:

$$g_i(x_i, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \; e^{-\frac{1}{2}(x_t-\mu_t)^T \Sigma_i^{-1}(x_t-\mu_i)} \quad (2)$$

With $d = 3$ represent the dimension of the RGB vector which is assumed to be independent random variables, while the covariance Σ_i is assumed to be a diagonal matrix for simplicity and low computation reason.

A mixture of components is initially built around a mean image, which is computed over a number of frames of the scene of static background. A Gaussian probability density function $f_i(x, m_i, \sigma_i)$ is associated to each pixel of the mean image as follows:

$$m_i = \frac{1}{t} \sum_t x_i\,(t) \quad (3)$$

$$\sigma_i = \frac{1}{t-1} \sum_{\mathrm{t}} (x_i(t) - m_i)(x_i(t) - m_i)^{\mathrm{T}} \quad (4)$$

Where $x_i(t)$ is the pixel value in frame t, m_i and σ_i are respectively the mean and covariance of pixel i.

By associating to the mean image the mixture density f of dimension n x m Gaussian components defined as follows:

$$p(x) = \sum_{i=1,n} \alpha_i f_i(x, m_i, \sigma_i) \quad (5)$$

With n and m represents the width and height of the image. $f_i(x, m_i, \sigma_i)$ is a Gaussian component with a mean and covariance given in equations (3) and (4), and α_i is the mixing weight equal to $\alpha_i = 1/\,nxm.$. By fitting all the components of the mixture f into a reduced and representative mixture g of N components, the SGGMM estimation is recast as a clustering problem so that the set of pixels of a region with similar colour are fitted to the same cluster and represented by the same Gaussian component.

An adaptive hierarchical clustering algorithm is used to create and update clusters. To determine the elements of the reduced mixture g closest to elements of the original mixture f, a distance minimization criterion between f and g is used. After setting the mixture components g_k, an observed pixel value is assigned to the component with the maximum posterior probability. Resulted on building a "support map" C_{map} to store the current component assignment for each pixel,

$$C_{map} = arg_{j=1,N} \max \{\log\,(p(x|g_i))\} \quad (6)$$

Where x is the pixel value. This support map obtained offline from the SGGMM model, will be used during the online image frame processing to perform segmentation.

The segmentation task is performed in each image frame to classify image regions as either background or foreground by evaluating the pixel likelihood, $\log\,(p(x|g_{cmap})$, in the new frame. If the likelihood is less than a user-defined threshold T_{seg}, the pixel is set to the foreground. Otherwise, it is set to the background.

III. THE ACOUSTIC SOURCE LOCALIZATION

In this work, interest will be given to loud sound detection. Thus, a focus will be given in estimating the location of the sound source, which is achieved by first detecting the burst of signal energy that exceeds the background acoustic energy and sending their corresponding times to a central unit for TDOA estimation and position calculation.

Under the assumption that the acoustic environment enables a reliable detection of these bursts of energy (by ignoring the effect of reverberation and multipath propagation), the problem of acoustic event detection can be written in the form of:

$$y_k = t_0 + \frac{1}{s}.\,\|P_k - X\| + e_k \quad (7)$$

With:

y_k :Time of signal detection;

s :The sound speed in the air (around 330m/s);

t_0 :The acoustic event triggering time;

X :The position acoustic source in world frame;

P_k :Position of the acoustic sensor k;

e_k :The detection errors;

The problem can be written as:

$$y = f(t, P_k, X) + e \quad (8)$$

A solution to this problem is possible using method of minimizing the sum square of errors in (11):

$$\hat{r}_n(X) = Arg\ Min_X\ \sum_n (y - f(t, P, X))^2 \quad (9)$$

With n represent the number of acoustic sensors pairs. Gauss Newton algorithm [16] iteratively finds the minimum of the sum of squares of residuals r_n. It works through guessing a random variable X^t as a possible initial solution and then trying to search for better solution that minimize this square of the residual by iterating equations (9) (10) and (11) until a minimal value is reached:

$$\boldsymbol{J}_r = \frac{\delta \boldsymbol{r}_i}{\delta \boldsymbol{X}_i}(\boldsymbol{X}_i{}^t) \quad (10)$$

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^{t} - (\boldsymbol{J}_r^T \boldsymbol{J}_r)^{-1} \boldsymbol{J}_r^T . \hat{\boldsymbol{r}}_n(\boldsymbol{X}^t) \qquad (11)$$

Errors in detection or due the non-convergence nature of the used numerical technique result in uncertainty of the estimated position. Thus, a more representative result is to be given by its statistics. Assuming that the estimated positions are of Gaussian random process, it can be considered as normally distributed random variables with a mean μ_X and a covariance Σ_X that can be estimate as follows:

$$\mu_X = E(X) \qquad (12)$$

$$\Sigma_X = E(\mu_x - X)^2 \qquad (13)$$

With *E(.)* represents the expectation operation.

The estimated acoustic source position is then projected on a duplicated copy of the captured image through transformation from world coordinates to image coordinates with, the exception that this duplicated copy is a greyscale image with intensities I_s ($0 \leq I_s \leq 1$) that interprets only the existence of acoustic sources so that regions of highest intensities represent higher probability of existing acoustic sources.

As the acoustic information arrives at slower rate than the video information, estimation of the acoustic source position is necessary. As the change in position of the active acoustic source is not noticeable during time frame $\Delta t = 1/Fr$ with Fr represents the camera frame rate. Therefore, the assumption that its position is fixed seems acceptable. For optimal estimation, the linear kalman filter (KF)[17] is used. It is defined by the following representation:

The state space S_{k+1} which is represented by:

$$S_{k+1} = \begin{bmatrix} x_{k+1} \\ y_{k+1} \\ Vx_{k+1} \\ Vy_{k+1} \end{bmatrix} \qquad (14)$$

The state estimation equation, and is given by:

$$S_{k+1} = F(S_k, v_k) \qquad (15)$$

With $\boldsymbol{v_k}$ is the process noise,

The transition model is given by:

$$F = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad (16)$$

While the observation model is given by the matrix y_k,:

$$y_k = H(S_k, n_k) \qquad (17)$$

With the measurement model *H* is given by:

$$H = \begin{bmatrix} x \\ y \end{bmatrix} \qquad (18)$$

The KF operates by propagating the mean and the covariance of the state through time using a set of equations described in [17] to estimate the state the system.

IV. AUGMENTING THE SGGMM USING ACOUSTIC INFORMATION

The SGGMM background model based on RGB color is augmented so that the the acoustic information I_s concatenated with the 3D feature vector [I_R I_G I_B]. This leads to a 4D vector [I_R I_G I_B I_s] that is called SGGMM based acoustic information.

The overall steps of the approach adopted in this work are illustrated in “Fig.1”:

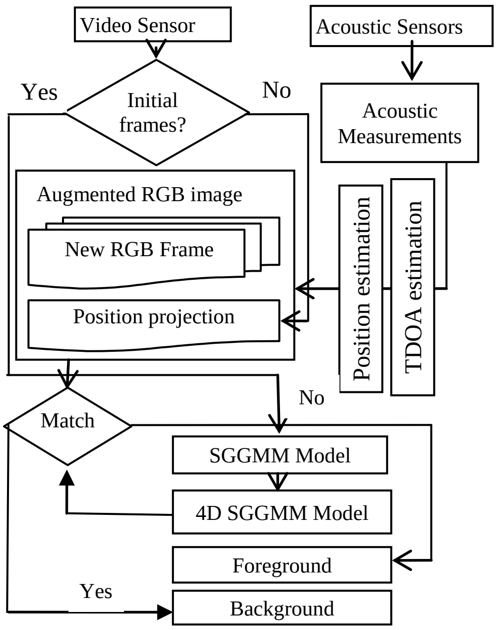


Fig. 1 different steps of the fusion approach

V. THE EXPERIMENTAL SETUP

To validate the proposed approach, experiments are done using a wireless sensor network composed of smart camera able to communicate with acoustic that are described in below:

A. The Vision Detection System

For vision detection, the CITRIC camera mote [18] is employed to implement the augmented SGGMM algorithm. It consists of a camera board attached to a Telosb wireless mote. The camera sensor is featured by a low voltage SXGA (1.3 mp). It supports image sizes SXGA (1280×1024), VGA, CIF, and any size scaling down from CIF to 40 × 30. The image array is capable of operating at up to 30 frames per second (fps) in VGA, CIF, and lower resolutions, and 15fps in SXGA.

The camera uses as the PXA270 processor[18], it has a maximum speed of up to 624MHz, 256KB of internal SRAM, and an MMX coprocessor to accelerate multimedia operations. Furthermore, it is connected to 64MB of SDRAM, 32 MB of NOR FLASH allowing the storage of the processed frames easily. Additionally, the PXA270 is a general purpose processor that runs embedded Linux and it is characterised by maturity of its software and development tools offering the opportunity to implement codes in C/C++. Communication among the camera components is insured via the CP2102 USB-to-UART[20] bridge controller used to connect the UART port of the PXA270 with a USB port on a personal computer for programming and data retrieval. The camera daughter board also has a JTAG interface for programming and debugging while the wireless communication is reserved for data flow from clients (acoustic sensors) to the server (the camera) with maximum data rate of 250 kbps per frequency channel.

B. The Acoustic Sensors

For acoustic measurement, MICAz motes [19] are used in this work. This type of motes is featured by its operation within the 2.4 GHz ISM band and is compliant with IEEE 802.15.4. It is designed especially for deeply embedded sensor network with capability of 250 kbps data rate. It is based on the low power microcontroller Atmel ATmega128L that runs TinyOS [20] from its internal flash memory of 10 kb. The processor board (MPR2400) of MICAz can be configured to run sensor application and radio communication simultaneously. Additionally, the motes can be connected to external peripherals such as sensors and data acquisition boards via 51-pin expansion connector. In this work, we used the MTS310 board [21] to benefit from its microphone circuit which is designed for audio applications.

C. Synchronisation and measurement of detection Time.

An application is developed in TinyOS (version 2.1.0) for measuring the loud sound detection time and its transmission. It is implemented at the acoustic sensors level. Since the used programming tool is deprived of any time synchronisation facility and to deal with this issue, we implement an algorithm on a coordinator mote that broadcasts a reference beacon to their neighbours insuring that all detector motes are within radio range of the coordinator. The implemented algorithm works through the following three steps:

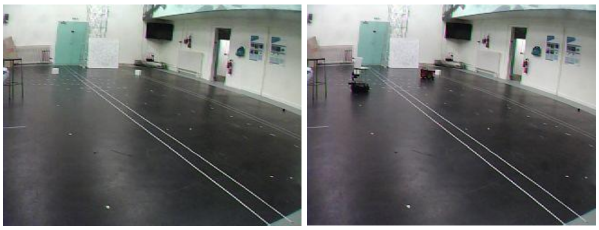
- Firstly, a coordinator sends a reference beacon to their neighbours using physical-layer,
- Secondly, group of five (5) motes situated at different locations but not far from target, receive the coordination message and prepare for the detection of a loud sound that exceeds a predefined threshold;
- Thirdly, the base station (the CITRIC camera mote with Telosb in this experiment) receives a set of detection times from the sensor motes.

In addition to calculating the TDOA and estimating the position of the acoustic source using the LS, the base station projects this positions in a duplicate image to the captured one as illustrated in section (III). The fusion of the two types of data using the augmented SGGMM approach is illustrated in section (IV).

VI. EXPERIMENTS AND RESULTS

The tests have been concluded in an indoor environment (Heaviside Laboratory at Cranfield University) and the testing was limited to images sequences with a resolution of 320×240 to ensure the shortest processing time.

In the first scenario “Fig. 2”, the acoustic events were triggered from the one in left of two moving robots “Fig. 2-b”. As a result, the two were detected successfully using SGGMM only “Fig.4” (first raw). While the silent robot is ignored using the augmented SGGMM “Fig. 5”, (first row).

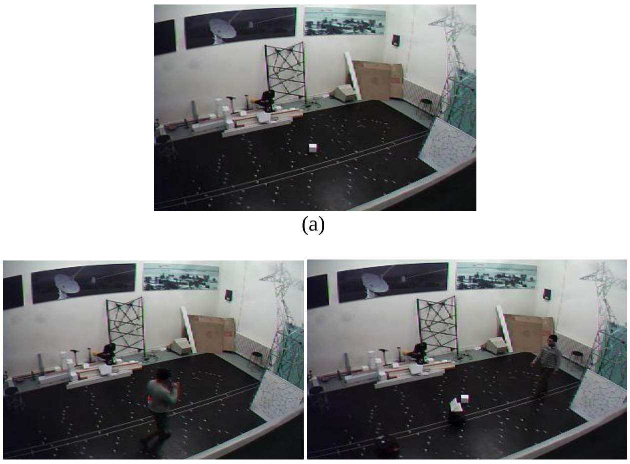


(a) (b)

Fig. 2 First scenario

In the second scenario “Fig. 3-b” the loud sound was caused by a walking man while in the third scenario “Fig. 3-c”, it was a robot (in the left down) that causes the sound. In the two cases the active moving sources are hardly distinguishable from the background using the SGGMM only, and this is because the similarity between the intensity of their colours is and the background colours “Fig.4” (second and third row).

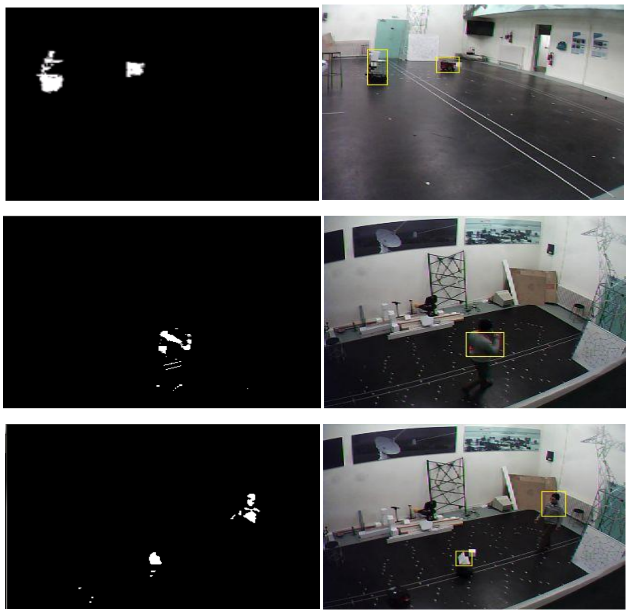
However, augmentation of the 3D RGB vector enabled improvement in detecting the full regions of the interested targets “Fig. 5” (second and third row).



(a)

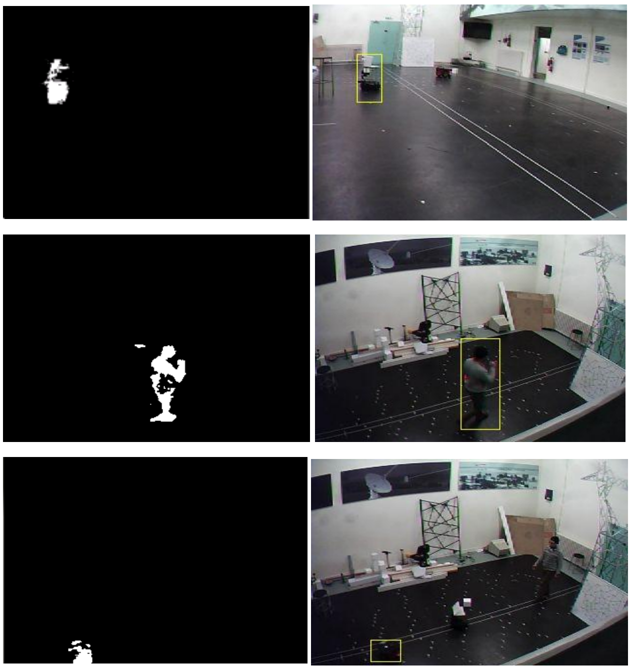
(b) (c)

Fig. 3 second and third scenario



(a) (b)

Fig. 4 results of SGGMM color background model: (a) segmentation using colour only based SGGMM background model. (b) detected targets.



(a) (b)

Fig. 5 results of the augmented SGGMM color background model (a) segmentation using colour combined with acoustic information based SGGMM background model, (b) detected active acoustic sources.

The average execution time of the different modules of the algorithm is illustrated in “Fig. 6”. Augmenting the 3D RGB vector using the acoustic information increases the execution time by around a quarter the time allocated for processing each frame using SGGMM only. By setting 320x240 as a resolution and with a speed of 624MHz for the camera processing unit, it reaches an overall processing time for about 650 ms per frame, a speed that enables the detection of moving active objects with moderate speed and can be improved using hardware with better specification.

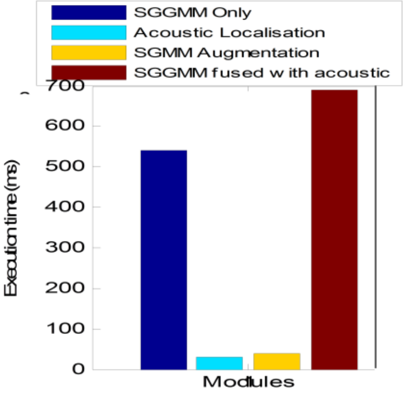


Fig. 6 repartition of computation time

VII. CONCLUSION

The problem of vision detection of moving objects enhanced by acoustics is investigated using an innovative Gaussian mixture models (GMM) approach in an embedded system. The feature vector of SGGMM used for background subtraction is augmented to include the acoustic information in the scene.

As matter of computational resource, the technique is not very demanding which enable the algorithm to be successfully implemented in WSN with components of similar computational capabilities. Results of the implementation have demonstrated that the proposed fusion allows a significant improvement in detecting the active acoustic sources. However, the adopted acoustic sensors suffer from limited processing capacity leading to relying only on pressure feature of the acoustic signal. Hence, the use of this technique for detection and localisation in complex scenarios is conditioned by the use of advanced acoustic sensing devices with high signal processing capabilities

REFERENCES

[1] S.Spors, R.Rabenstein, N.Strobel, "*A Multi-sensor Object Localization System*," Proceedings of the Vision Modeling and Visualization Conference, Stuttgart, Germany, Nov 2001, pp. 19-26.

[2] G.Friedland, C.Yeo, and H.Hung. "*Visaul speaker localization aided by acoustic models*". In proceedings of ACM multimedia, October 2009.

[3] Z.Li;Herfet.T,Grochulla.M. and Thormahlen.T “*Multiple active speaker localization based on audio-visual fusion in two stages*”, In Proc of IEEE Conference on multisensor Fusion and Integration for Intelligent Systems (MFI), 2012.

[4] A.C.Sankaranayanan, Q.Zheng, R.Chellappa, V.Cevher, J.H.McClellan, and G.Qian.”*Vehicle tracking using acoustic and video sensors*”. IEEE transaction on Acoustics, Speech and Signal Processing-2004, Volume: 3, Pg: 793-6.

[5] B.Bunin, A.Sutin, G.Kamberov, H.S.Roh, B.Luczynski, M.Burlick. “*Fusion of acoustic measurements with video surveillance for estuarine threat detection*”. In Proc of SPIE 6945 (Optics and Photonics in Global Homeland Security IV, 694514-2008).

[6] M.Piccardi., “*Background subtraction techniques: a review*”. IEEE International Conference on Systems, Man and Cybernetics, 4:3099-3104, October 2004.

[7] Z.Yi ; F.Liangzhong. “*Moving object detection based on running average background and temporal difference*”. International Conference on Intelligent Systems and Knowledge Engineering (ISKE), pages 270-272, November 2010.

[8] W. Grimson, and C.Stauffer. “*Adaptive background mixture models for real time tracking*”. IComputer Vision and Pattern Recognition Conference (CVPR), 1999.

[9] Z.Yi and F.Liangzhong.”*Real-time tracking of non-rigid objects using mean shift*”. Computer Vision and Pattern Recognition, 2:142-149, June-2000.

[10] D. Comaniciu B. Han and L.S. Davis. “*Sequential kemel density approximation through mode propagation: applications to background modeling*”. In Proc of. Asian Conf. on Computer Vision, pages 270- 272, Jan 2004.

[11] M.S.Kemouche, N.A*ouf, "A Gaussian mixture based optical flow modeling for object detection* ," In proceeding of: Crime Detection and Prevention (ICDP 2009), 3rd International Conference.

[12] C.meesokho,”*Robust acoustic source locilisation in WSN*”, Phd thesis, University of Southern Calorina, 2007.

[13] J. C. Chen, R. E. Hudson, and K. Yao, "*Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field*" IEEE Transactions on Signal Processing, vol. 50, no. 8, pp. 1843-1854, August 2002.

[14] K. W. Cheung , H. C. So, W. K. Ma and Y. T. Chan "*Least squares algorithms for time-of-arrival-based mobile location*", IEEE Trans. Signal Processing, vol. 52, pp.1121 -1128 2004 .

[15] Gene T. Whipps, Lance M. Kaplan, and Raju Damarla. *Analysis of Sniper Localization for Mobile, Asynchronous Sensors*. In Proc of SPIE volume 7336, May 11, 2009.

[16] S. Gratton, A.S. Lawless and N.K. Nichols, “*Approximate Gauss-Newton methods for nonlinear least squares problems*”. Numerical Analysis Report 9/04, University of Reading.

[17] Welch G, Bishop G. *An Introduction to the Kalman Filter*. SIGGRAPH, 2001.

[18] P.Chen, P.Ahammad, C.Boyer, S.Huang, L.Lin, E.Lobaton, and others ”*citric: a low-bandwidth wireless camera network platform. International Conference on Distributed Smart Cameras*”, 2008. Pg:1-10-Stanford,CA.

[19] M.Drieberg, N.A.Ali and P.Sebastian. “*deployment of Micaz mote for wireless sensor network application*”. International conference on computer applications and industrial electronics 2011 (iccaie 2011).

[20] D.Gay P.Levis. “*Tinyos* programming”. Cambridge University Press, 2009.

[21] Crossbow team. *“Mts/mda sensor board user’s manual”*. 2007.

HybridLog: an Efficient Hybrid-Mapped Flash Translation Layer for Modern NAND Flash Memory

Mong-Ling Chiao and Da-Wei Chang

***Abstract*—**A Flash Translation Layer (FTL) emulates a block device interface on top of flash memory for supporting traditional disk-based file systems. Because of the erase-before-write feature of flash memory, out-of-place update is usually adopted in an FTL and a cleaning procedure is typically used to reclaim obsolete data. Hybrid-mapped FTLs are widely adopted in flash memory storage devices such as secure digital memory cards (SDs) and USB flash disks (UFDs). However, many traditional hybrid-mapped FTLs have limited support to modern NAND flash memories and could have high cleaning cost in the face of random writes. In this paper, an efficient hybrid-mapped FTL supporting modern NAND flash memories is proposed. Modern flash memory support is achieved by enabling log-style write in all the blocks and efficient use of spare area. The use of log-style write also achieves high efficiency by eliminating writes of dummy pages to the data blocks and by reducing the write traffic to the small-sized log area due to page collisions. Results from six realistic and benchmark based workloads show that the proposed FTL outperforms existing hybrid-mapped FTLs by up to 17.8 times in terms of cleaning cost.

***Keywords*—**flash translation layer, NAND flash memory, performance, storage management

I. Introduction

NAND flash memory is widely applied in computer and consumer electronic devices due to its small size, shock resistance, non-volatility and low power consumption. A NAND flash module is composed of a number of blocks, each of which in turn consists of a number of pages, and read/write operations are performed in units of one page. Each page is typically made up of a 512-byte to 2-Kbyte main area used to store user data and a 16-byte to 64-byte spare area used to store metadata such as page mapping information and error correction codes (ECC). A software component called Flash Translation Layer (FTL) is usually used to emulate a block device on top of the flash memory to support traditional disk-based file systems.

In contrast to magnetic disk, flash memory pages cannot be overwritten before being erased, and erase operations are performed in units of an entire block. The number of erase operations that can be done on a specific block is limited (usually between 300 and 100,000). To avoid erasing entire block for each logical page overwrite, an FTL directs each page overwrite to a free physical page. The page containing the stale data is then reclaimed by a cleaning procedure, and cleaning cost is a key factor to the performance of an FTL.

With the development of flash memory, new restrictions are imposed on flash memory chips, and an FTL should follow these new restrictions so as to be applied on these modern chips. Specifically, a new programming (i.e., write) restriction called *consecutive programming* is imposed on most modern flash memories [1], whereby pages have to be programmed in consecutive order (i.e., from lower-numbered pages towards higher-numbered pages) within a block. Moreover, Multiple Level Cell (MLC) NAND achieves lower cost by allowing multiple bits to be stored in a single cell. However, compared to Single Level Cell (SLC), MLC has a higher bit error rate and thus requires stronger ECC, which consumes more spare area space, preventing FTLs requiring large space of the spare area from being applied on it.

An FTL determines the physical location of each logical page and manages the mapping between the logical page numbers (LPNs) and the physical page numbers (PPNs). The mapping can be done at two different granularities: page-level and block-level. Page-mapped FTLs [2], [3] allow each physical page to be associated with any logical page. Such flexible mapping leads to low cleaning cost when the storage utilization of the flash storage is not high. For a large NAND flash memory, such a fine-grained address translation scheme requires either a large memory space (in the case of in-RAM mapping table) or frequent flash memory accesses (in the case of in-flash mapping table) to maintain the most up-to-date mapping table since each logical page has a corresponding entry in the table.

Block-mapped FTLs [4], [5] use coarse-grained mapping to achieve lower RAM consumption for the mapping information. In a block-mapped FTL, each logical block has an associated data block to accommodate page writes to that logical block. These FTLs require each logical page to be written only to its corresponding offset of a physical block, leading to poor

This work was supported in part by the National Science Council, Taiwan, Republic of China, under Grant NSC 101-2221-E-006-098-MY3 and Grant NSC 101-2221-E-006-150-MY3.

M. L. Chiao is with the Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan 300, R.O.C. (e-mail: jackciao.cs94g@nctu.edu.tw).

D. W. Chang is with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan 701, R.O.C. (phone: 886-6-2757575 ext. 62551; fax: 886-6-2747076; e-mail: davidchang@csie.ncku.edu.tw).

performance and preventing efficiently support of consecutive programming. Hybrid-mapped FTLs [6]–[12] manage most of the blocks (i.e., data blocks) via the block-level mapping approach. By storing hot pages (i.e., frequently-updated pages) in a small number of log blocks, which are managed by the page-level mapping approach, hybrid-mapped FTLs achieve performance superior to block-mapped FTLs.

Currently, hybrid-mapped FTLs are widely used in flash memory storage devices such as secure digital memory cards (SDs) and USB flash disks (UFDs). These storage devices are often used for backup or exchange photo, audio, video and document files. Since the sizes of these files are typically large, accessing these files usually induces sequential read/write transactions on the storage devices. In a hybrid-mapped FTL, sequential read/write transactions can usually be satisfied with low cleaning cost. However, random read/write transactions still exist in SDs or UFDs due to accesses to file system metadata, and a hybrid-mapped FTL could have high cleaning cost due to these accesses. Moreover, many hybrid-mapped FTLs cannot support consecutive programming efficiently since the constraint of the block-level mapping (i.e., each logical page can only be written to its corresponding offset in a physical block) is still valid for the data blocks.

In this paper, a novel hybrid-mapped FTL called HybridLog is proposed to support modern NAND flash memory and to achieve low cleaning cost. To allow consecutive programming while reducing cleaning cost, HybridLog enables log-style writes to all the blocks (including the data blocks) in the flash memory. To support log-style writes to all blocks, intra-block mapping information is stored in the spare area of each written page. Since only a small space is required in the spare area for the mapping information, many modern SLC/MLC flash memories can be supported. The performance of HybridLog and two well-known FTLs are compared under various realistic and benchmark-based workloads. The performance results show that, HybridLog outperforms these two hybrid-mapped FTLs by up to 17.8 times in terms of cleaning cost The write amplification ratio can be reduced by up to 58%.

II. Background and Related Work

A. Background and Terminology

An FTL provides a method to allow a NAND flash memory device to emulate a random access block device efficiently. Since a programmed NAND flash page need to be erased before it is programmed again, updating data in place is inefficient because it not only takes a time-consuming erase operation, but also incurs the wear-leveling issue [13]-[15]. Thus, most FTLs use an out-of-place update mechanism to update a logical page. In the out-of-place mechanism, each NAND flash memory page is in one of the three states, *free*, *live* and *dead*. A page becomes free after being erased. A free page can be used to accommodate page writes, and it becomes live after being written. After the new data have been written to another free page, the live page containing the old data becomes dead. Dead pages can be reclaimed by a *cleaning* procedure, which selects victim blocks according to a cleaning policy, copies the live pages of the victim blocks to free pages of other blocks, and erases the victim blocks. Cleaning is time-consuming since it involves live page copying and block erasure. As a consequence, the cost of cleaning is a key factor to the performance of an FTL.

In this paper, two metrics related to the cost of cleaning are used to measure the performance of an FTL. The first one is the cleaning cost, which is defined as the time spent on the cleaning procedure during workload execution. The second one is the Write Amplification Ratio (WAR), which is defined as

$$WAR = (W + C) / W \tag{1}$$

In (1), W and C represent time for serving user write requests and the time for cleaning (i.e. cleaning cost) during the execution of the workload, respectively. The ratio 1.5 means that the time spent on cleaning is half of the time for serving user writes in the given workload.

B. Flash Translation Layer

The mapping between LPNs and PPNs can be done at page level or block level. Page-mapped FTLs directly translate each LPN to a PPN and use the out-of-place update mechanism to handle page overwrites. The mapping is flexible since each physical block can accommodate pages belonging to any logical blocks. However, the mapping information is traditionally stored in RAM and such fine-grained mapping requires a large RAM space for a large sized flash memory. Recently, several RAM-space-efficient page-mapped FTLs address this problem by storing the mapping information in the flash memory and caching the recently used information in RAM [2]. Cleaning in page-mapped FTLs is done by reclaiming blocks (i.e., copying live pages in victim blocks to blocks with free pages and then erasing victim blocks). After the reclamation, the erased blocks can be used to accommodate future writes.

After a page-mapped FTL has selected a victim block for cleaning, it has to identify the live pages of the victim block. Querying/updating the mapping information of the live pages is needed after cleaning. If the FTL stores the page state and mapping information of each page in RAM, query/update of the mapping information can just be done in RAM. However, as mentioned above, a large RAM space would be required for a large sized flash memory. In memory-constrained consumer storages such as SDs or UFDs, mapping information of a page-mapped FTL can only be stored in the flash memory (and cached in RAM). Thus, after a victim block has been selected, extra flash memory read/write operations need to be performed to identify the live pages in the victim block and to locate the physical locations of the mapping information of the live pages. Note that, victim blocks are cold blocks and thus their information is seldom cached in RAM. If there are many live pages in the victim block (i.e., high storage utilization) and the mapping information of the live pages are stored in many different mapping pages, many flash memory reads/writes are required for querying and updating the mapping information.

Block-mapped FTLs achieve lower RAM consumption for the mapping information by using coarse-grained mapping. In a block-mapped FTL, each logical block has an associated data block to accommodate page writes to that logical block. Given a page write, the LPN is divided by the number of pages in a block to obtain the logical block number (i.e., the quotient) and the page offset (i.e., the remainder). The former is used to index the mapping table to obtain the physical address of the data block, and the latter is used to locate the target page in the data block. If the target page is live (i.e., *page collision*), in-place update is used. Besides, if the target page is not consecutive to the last written page of the data block, dummy pages has to be written between the last written page and the target page (i.e., *page padding*).

The block-mapped approach requires each logical page to be written to a fixed offset of a data block, increasing the frequency of block reclamation. In addition, it also prevents efficient support of consecutive programming due to the writes of extra pages (i.e., dummy pages).

Several hybrid-mapped FTLs have been proposed to achieve performance superior to block-mapped FTLs, while retaining the small size of the mapping information. In these FTLs, most of the blocks (i.e., data blocks) are managed via the block-level mapping approach. However, by managing a few log blocks via the page-level mapping approach to accommodate frequently-updated pages, the hybrid-mapped FTLs reduce the frequency of data block erasure. Hybrid-mapped FTLs utilize the out-of-place update mechanism. Page writes that cannot be accommodated by the data blocks are satisfied by the log blocks, and the pages containing the old data become dead. Cleaning in hybrid-mapped FTLs is done by reclaiming log blocks (i.e., merging log blocks with their associated data blocks). After the reclamation, free log blocks are obtained to accommodate future writes.

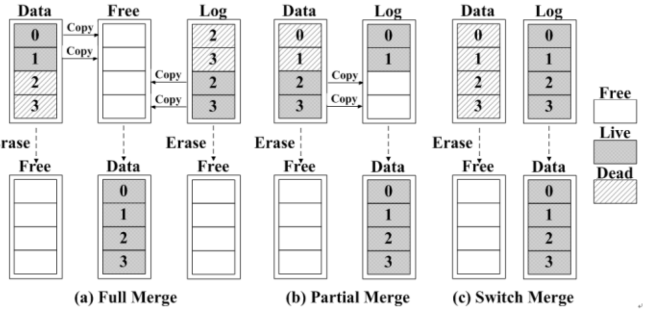


Fig. 1 three types of merge operations

As shown in Fig. 1, three types of merge could occur depending on the status of the log block and the associated data blocks. In Fig. 1(a), the merge operation copies the live pages from the data and the log blocks to a free block *F*. After the copying, the old data and log blocks are both erased and the block *F* becomes the new data block. This is called *full merge*. In Fig. 1(b), the merge can be done by copying the live pages in the data block to the free space of the log block, erasing the data block, and finally prompting the log block as the new data block. This is called *partial merge*. In Fig. 1(c), all the up-to-date data were written in the log block sequentially and thus the merge operation can be done simply by switching the roles of the log and data blocks and erasing the original data block, which is called *switch merge*. Of the three types of merge operations, the switch merge has the lowest cost while the full merge results in the highest cost.

Many hybrid-mapped FTLs cannot support consecutive programming efficiently since the constraint of the block-level mapping (i.e., each logical page can only be written to its corresponding offset in a physical block) is still valid for the data blocks. In these hybrid-mapped FTLs, therefore, dummy page writes may still be required during workload execution. Although FTLs such as Superblock [8] avoid this problem, they either have limited support to large-block MLC flash memory, due to the storing of a large amount of information in the spare area and thus prohibiting the use of strong ECC, or suffer from inferior performance. In this paper, a hybrid-mapped FTL supporting modern NAND flash memory and achieving performance superior to existing hybrid-mapped FTLs is proposed.

III. DESIGN OF HYBRIDLOG

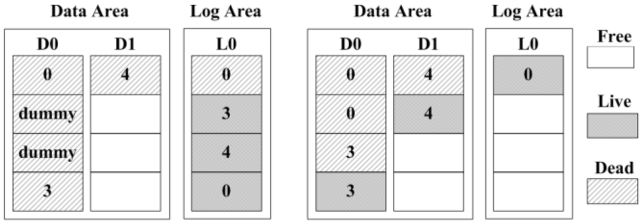
The same as traditional hybrid-mapped FTLs, HybridLog divides the flash memory into two areas, a large data area, containing data blocks managed by block-level mapping, and a small log area, containing log blocks managed by page-level mapping. Each logical block has an associated data block to accommodate writes to that logical block, and thus the user perceived storage size is the data area size. However, HybridLog adopts a novel architecture to allow consecutive programming and to reduce cleaning cost. The details of HybridLog are described below.

A. Architecture of HybridLog

Different from traditional hybrid-mapped architecture, the HybridLog architecture enables log-style writes to all the blocks in the flash memory, especially the data blocks, allowing consecutive programming. Since data blocks are written in a log order, log blocks are used only after a data block is full. The log-style writes keep 100% utilization of the data blocks even under the random-write workloads. Fig. 2 illustrates an example showing the difference between HybridLog and the traditional hybrid-mapped architecture. Assume that the flash memory consists two data blocks and one log block, with each block containing four pages. Fig. 2(a) and Fig. 2(b) illustrate the result of the page write sequence (0, 0, 3, 4, 3, 4, 0) under the traditional hybrid-mapped and HybridLog architectures, respectively.

In Fig. 2(a), although the first write to (logical) page 0 can be served by D0, the second write to the page 0 has to be served by the log block due to page collision. Moreover, the first write to page 3 has to be served by the last physical page of D0 due to the use of block mapping in the data area, and two dummy pages have to be written to the second and third pages of D0 before the write of page 3 to follow consecutive programming. Such dummy page writes increase not only the write response time but also the WAR. The second writes to logical pages 3 and 4 also incur page collisions in D0 and D1, respectively, and therefore these two page writes have to be satisfied by the log block. After the third write to page 0, the log block is full and cannot accommodate further page writes. In Fig. 2(b), except the last page write to page 0, all page writes are proceeded in log-style in the corresponding data blocks D0 and D1. The last write to page 0 is satisfied by the log block because the corresponding data block D0 is full. After the last logical page is written, the log block still has three free pages to accommodate further page writes.

As a result, HybridLog not only eliminates unnecessary dummy page writes but also reduces the write traffic to the small-sized log area. Therefore, cleaning cost and WAR can be reduced. In the following, the technique to enable log-style writes in all blocks is described.



(a) Traditional hybrid-mapped architecture

(b) HybridLog architecture

Fig. 2 traditional hybrid-mapped architecture vs. HybridLog architecture

B. Log-Style Writes

HybridLog uses the block-mapped approach to manage the data blocks. If the target page that needs to be written is not consecutive to the last written page in the data block, traditional block-mapped and hybrid-mapped approaches fill dummy content to satisfy the consecutive programming restriction of modern NAND flash memory. This causes overhead in both time and flash memory space, and the space overhead could be large for small random writes.

Although log-style writes in a data block can be achieved by using page level mapping for data blocks. This causes large RAM space for the mapping information. To enable log-style writes in a data block while preventing increasing the RAM size for the total mapping information, HybridLog stores the intra-block mapping information (i.e., the *physical page offset* of each logical page in a data block) in the spare area of each written page. The information is organized as a two-level mapping table. As shown in Fig. 3, which assumes 64-page blocks, the first-level mapping is called the Mapping Directory (MD). Each entry in the MD refers to a Mapping Table (MT), and each entry in the MT records the physical page offset for the corresponding logical page. It can be regarded as each logical block being divided into a number of groups, with each group containing a fixed number of contiguous logical pages. Each MT records the mapping information of a group and the MD keeps track of the location of all the MTs in the logical block. In Fig. 3, the block is divided into 2 groups with each group containing 32 contiguous logical pages. For each page write to a data block, the up-to-date MD and MT, derived from the information in the spare area of the last written page in that data block and the information of the to-be-written page, are stored in the spare area of the to-be-written page.

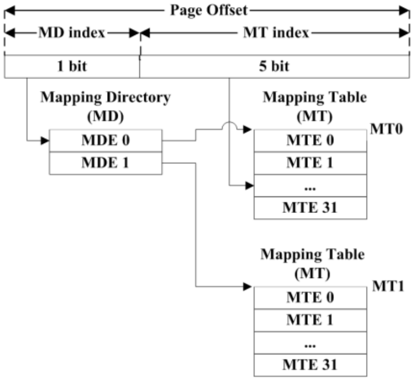


Fig. 3 two-level intra-block mapping

Fig. 4 illustrates the spare area format of each written page, which is divided into 3 sections: data information (DI), MD and MT. The DI contains bad block indicator, LPN and ECC, while the other sections are used for recording the mapping information. Assume *B* and *G* denote the number of pages per block and the number of pages per group, respectively, a *B/G*-entry MD and a *G*-entry MT are included in each spare area. Each entry has a size of $\log_2 B$ bits since it is used to locate a page in the physical block. The size requirement of the spare area space will be analyzed later.

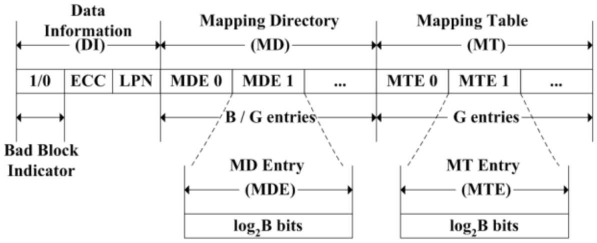


Fig. 4 spare area format of each written page

Fig. 5 shows the steps of writing a page with LPN 874 to the corresponding data block, assuming 64-page blocks and 32-page groups. First, the LPN is divided into logical block number (LBN) 13 and page offset 42, and the latter is in turn divided into MD index 1 and MT index 10, meaning the page offset is stored in entry 10 of MT1. Second, the LBN is used to index the block-level mapping table to obtain the physical block number 7. In that physical block, the first free page will be used to accommodate the write. From the old state of physical block 7 shown in the top right of Fig. 5, page 3 is the first free page of physical block 7, and thus it is used to accommodate the write. Third, the location of the MT1 is obtained by indexing the MD of page 2, the last written page in the physical block. From the figure, the entry 1 of MD refers to page 1, indicating that MT1 is stored in the spare area of page 1. Moreover, the entry 10 of MT1 also refers to page 1, meaning the old data of logical page

874 is stored in page 1. Fourth, entry 1 of MD and entry 10 of MT1 are both updated to refer to page 3, and the data together with the latest mapping information are written to that page. Finally, the page 1, which stores the old data of the logical page, is marked as dead, and page 3 is marked as live.

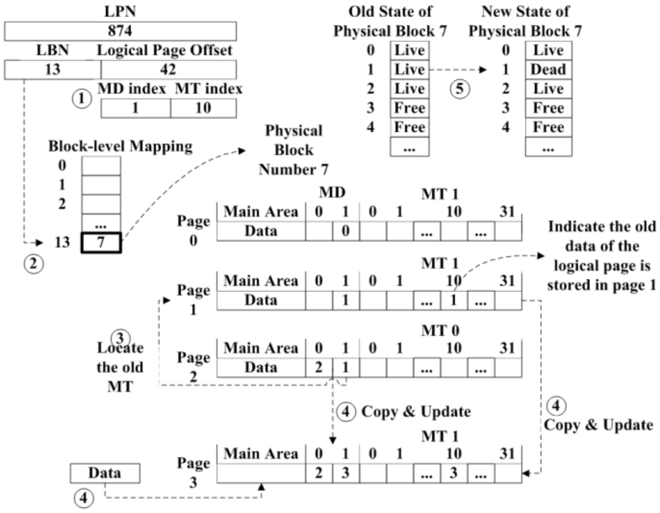


Fig. 5 steps of writing a page

From this example, the content of the new MD and MT (i.e., the MD and MT in page 3) are obtained from the old MD and the old MT, respectively. Specifically, assuming logical page l is to be written to physical page p of the target data block, and $MD(i)$ and $MD'(i)$ denote the ith entry of the old and the new MDs, respectively, the MD and MD' can be obtained by (2) and (3), respectively.

$$MD = \begin{cases} \text{NULL, if } p = 0, \\ MD \text{ in (thesparearea of) page } p-1, \text{otherwise.} \end{cases} \quad (2)$$

$$MD'(i) = \begin{cases} p, \text{ if } i = \lfloor l/G \rfloor, \\ MD(i), \text{ if } (i \neq \lfloor l/G \rfloor) \text{ and } (MD \neq \text{NULL}), \\ \text{NULL, otherwise.} \end{cases} \quad (3)$$

From (2), the old MD is stored in the last written page (i.e., page p-1) of the target block. Generally, all the entries of the new MD except from the one that refers to the MT containing the logical page l are copied from the corresponding entries of the old MD. However, in the case of a first-page write to a physical block (i.e., p = 0), the old MD does not exist. In that case, all the entries of the new MD are set as NULL, except from the entry that refers to the MT containing the logical page l. Similarly, assuming $MT(i)$ and $MT'(i)$ denote the ith entry of the old and the new MTs, respectively, the MT and MT' can be obtained by (4) and (5), respectively.

$$MT = \begin{cases} \text{NULL, if } MD(\lfloor l/G \rfloor) = \text{NULL}, \\ MT \text{ in page } k \text{ where } MD(\lfloor l/G \rfloor) = k, \text{otherwise.} \end{cases} \quad (4)$$

$$MT'(i) = \begin{cases} p, \text{ if } i = l, \\ MT(i), \text{ if } (i \neq l) \text{ and } (MT \neq \text{NULL}), \\ \text{NULL, otherwise.} \end{cases} \quad (5)$$

From the above description, it can be seen that page read/write requires additional spare area reads to lookup the intra-block mapping. To reduce the spare area reads, recently used intra-block mapping is cached in RAM. Each cache entry stores the mapping of a data block (i.e., the up-to-date MD and the associated MTs). Due to the temporal and spatial locality of page access, few cache entries are adequate for achieving a high cache hit ratio.

C. *Space Area Requirement of HybridLog*

In the following, the spare area space requirement is analyzed. According to Fig. 4, the space required by the intra-block mapping M_{spare_area} can be expressed in (6). To allow the mapping to be fitted into the spare area, (7) should hold, given S and D denoting the sizes of the spare area and DI, respectively.

$$M_{spare\_area} = G * \log_2 B + (B/G) * \log_2 B \quad (6)$$

$$M_{spare\_area} \leq S - D \Rightarrow [G + (B/G)] * \log_2 B \leq S - D \quad (7)$$

From (7), a set of possible values of G (i.e., pages per group) can be obtained for given values of B, S and D. Table I shows the common modern flash memory configurations and the corresponding possible values of G. Typically, the values of B (i.e., pages per block) are 64 and 128 for SLC and MLC, respectively. The value of S (i.e., spare area sizes) is typically 64 bytes for both SLC and MLC. The value of D is equal to the size of ECC plus the sizes of LPN (typically 4 bytes) and bad block indicator (typically 1 byte). In general, the number of bits required by the ECC depends on the flash type and the error correction algorithm. Most SLC and MLC modules require correcting 1-bit and 4-bit errors for each 512 bytes of data, respectively, and ECC sizes in this table are calculated based on satisfying that requirement by using the BCH algorithm [16], the most widely-used error correction algorithm in flash storage devices.

Although the Superblock FTL also stores the mapping information in the spare area, it tightly limits the maximum number of physical blocks allocated to a logical block (i.e., 8 blocks). This could lead to high cleaning cost due to the use of small buffers to accommodate page updates of frequently-updated logical blocks. Moreover, Superblock consumes a larger amount of spare area space due to the recording of multiple physical block numbers, prohibiting its use on some types of MLC flash. For example, in Table I, for the MLC NAND flash with main/spare area size as 2048/64 bytes, Superblock leaves only 8 bytes in the spare area for the ECC and thus it cannot be used on that type of flash memory. Decreasing the maximum number of physical blocks allocated to a logical block allows the support of more types of MLC flash with the cost of degraded performance. Although Superblock FTL can adopt an alternative approach, which stores the mapping information in the user area of dedicated pages called *map pages*, it incurs extra programming overhead for these map pages. In contrast, HybridLog uses the page-mapped approach in the log area and thus the entire log area can be used to buffer page updates of any logical blocks. In addition, it supports more MLC flash memories since only page offset information is stored in the spare areas.

TABLE I Common Flash Memory Configurations and the Corresponding Group Sizes

Flash Type	**SLC**	**MLC**
Pages Per Block	64	128
Main/Spare Area Size	2048/64 bytes	2048/64 bytes
ECC Size	7 bytes	26 bytes
Pages Per Group	1, 2, 4, 8, 16, 32, 64	8, 16

IV. Performance Evaluation

A trace-driven simulator was developed for the performance evaluation. In addition to HybridLog, two well-known hybrid-mapped FTLs, FAST and Superblock, were also implemented in the simulator for performance comparison. The simulation results show that HybridLog has superior performance to the other hybrid-mapped FTLs.

A. *Experimental Setup and Traces*

TABLE II Default Values of the Simulation Parameters

Parameters	**Default Values**
Storage Size	80 Gbytes (655,360 blocks)
Log Area Size	16,384 blocks (2.5 % of the storage)
Number of Pages Per Block	64
Number of Pages Per Group	32
Page Size (Main + Spare Area)	2112 (2048 + 64) bytes
Block Erase Time	2000 us
Page Read Time	88 us
Page Write Time	263 us
Number of Cache Entries	8

TABLE III Traces

Traces	**Sectors Written**	**Ave. Write Size**
LinuxPC	107,111,668	71.7
PostMark	8,816,216	6.6
LargeFile	60,149,736	748.6
Fin1	30,517,409	7.4
Fin2	3,810,800	5.8
4VMs	109,804,512	35.3

Table II shows the default values of the simulation parameters. All the time-related values are obtained from the specification of the modern MLC NAND flash chip, as in [10]. Six traces gathered from the execution of real workloads or benchmarks are used for performance comparison, also shown in Table III.

All the workloads are the same as in [11], [12]. The *LinuxPC* trace is a 10-day real workload of daily user activities such as web browsing, file editing, multimedia playing and program compilation, on Linux environment. The *Postmark* trace was gathered from the execution of the *PostMark* file system benchmark, which first creates 80,000 small files, and then performs 1,000,000 transactions such as create, delete, read, and append on the files. The *LargeFile* trace is the workload of creating and deleting MP3 files, whose average size is around 4 Mbytes, and is dominated by large sequential writes. The ratio of file creation to deletion is set as 10 and the workload terminates until the total number of existing files exceeds 10,000. The *Fin1* and *Fin2* traces are workloads of OLTP applications. The *4VMs* trace is a mixed workload generated from the execution of 4 virtual machines on top of a hypervisor. Each virtual machine, equipped with 768-Mbyte memory and 20-Gbyte virtual disk, runs one of the following workloads on the Linux kernel 2.6.31: file server, web proxy, mail server and OLTP, obtained from the *FileBench* file system benchmark.

In the following, the effect of log-style write is first presented. Then, an overall performance comparison among different FTLs is shown.

B. *Effect of Log-Style Write*

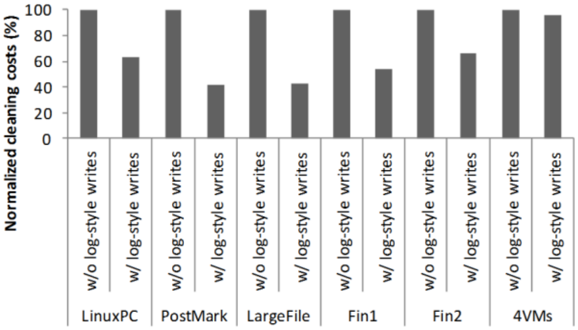


Fig. 6 normalized cleaning cost with and without log-style writes

To evaluate the performance of log-style write, we measure the cleaning cost with and without the presence of log-style write in the HybridLog FTL. Since the values of the traces have different orders of magnitude, they are normalized to the cleaning cost without log-style write. As shown in Fig. 6, using log-style write is effective in five of the six workloads, i.e., *LinuxPC*, *PostMark*, *LargeFile*, *Fin1* and *Fin2*. In these workloads, using log-style write can reduces the cleaning cost by up to 58%.

The reason can be seen in Fig. 7 and Fig. 8. Fig. 7 shows the average number of dummy pages that have been written when a data block becomes full. As mentioned before, dummy pages have to be written in each page padding operation to follow consecutive programming. From the figure, about 4 to 16 dummy pages in average were written in a data block. This wastes the flash memory space since these dummy pages do not accommodate any new user data. Moreover, further page writes to the space occupied by the dummy pages cause page collisions and thus have to be satisfied by the log area. With log-style write in HybridLog, dummy page writes can totally be eliminated.

Fig. 8 shows the average number of page collisions in a data block with free pages. Without log-style write, the collided pages have to be written to the log area even in the case that the data blocks still have free space to accommodate the collided pages. With log-style write in HybridLog, the collided pages can be accommodated by data blocks if there is free space in the data blocks. This reduces the write traffic to the log area and thus leading to lower cleaning cost.

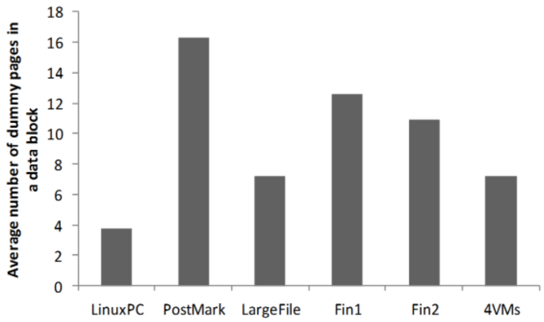


Fig. 7 average number of dummy pages in a data block

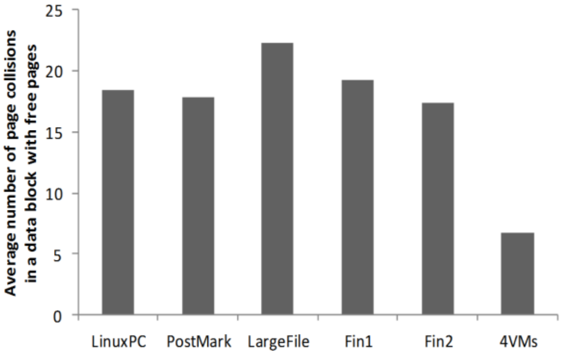


Fig. 8 average number of page collisions in a data block with free pages

In summary, as mentioned in Section III, log-style write allows every page of the data block to accommodate meaningful user data. Moreover, it reduces the write traffic to the small-sized log area and hence delays cleaning due to the fullness of the log area, resulting in a lower cleaning cost.

C. Overall Performance

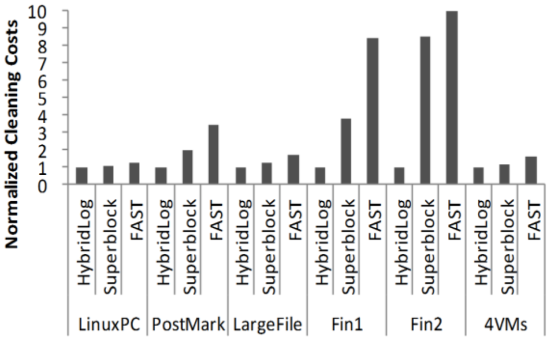


Fig. 9 normalized cleaning cost of Superblock, FAST and HybridLog

In this section, the overall performance of FAST, Superblock and HybridLog is compared. Fig. 9 shows the cleaning cost of the three FTLs under each trace. The results are normalized to the cleaning cost of HybridLog. From the figure, HybridLog outperforms FAST and Superblock by 30% (under *LinuxPC*) to 17.8 times (under *Fin2*) and 10% (under *LinuxPC*) to 7.5 times (under *Fin2*), respectively. Due to the reduction in the cleaning cost, HybridLog reduces the write amplification ratio (WAR), as defined in (1), by up to 1.73 and 0.65 when compared to FAST and Superblock, respectively.

V. Conclusions

In this paper, a novel hybrid-mapped FTL called HybridLog is proposed to support modern NAND flash memory and to achieve low cleaning cost. To allow consecutive programming required by modern NAND flash memories, log-style write is used for all the blocks in the flash memory, including the data blocks. To support log-style write to all the blocks, intra-block mapping information is stored in the spare area of each written page. Since only a small space is required in the spare area for the mapping information, many modern SLC/MLC flash memories can be supported. In addition to allow consecutive programming, log-style write to data blocks also eliminate writes of dummy pages to the data blocks and reduce the write traffic to the small-sized log area due to page collisions, which are both helpful for reducing the cleaning cost.

Through trace-driven simulation on six real or benchmark-based workloads, the effectiveness of log-style write and the superior performance of HybridLog compared to the other hybrid-mapped FTLs have been demonstrated. Specifically, HybridLog outperforms existing hybrid-mapped FTLs by up to 17.8 times in terms of cleaning cost and reduces the WAR by up to 1.73.

References

[1] *Open NAND Flash Interface 3.2 Specification*, Open NAND Flash Interface, 2013.

[2] A. Gupta, Y. Kim, and B. Urgaonkar, “DFTL: a flash translation layer employing demand-based selective caching of page-level address mappings,” *in Proc. Int. Conf. Architectural Support for Programming Languages and Operating Systems*, Washington, D.C., 2009, pp. 229–240.

[3] Y. G. Lee, D. Jung, D. Kang, and J. S. Kim, "μ -FTL: a memory efficient flash translation layer supporting multiple mapping granularities," *in Proc. ACM/IEEE Int. Conf. Embedded Software*, Atlanta, 2008, pp. 21–30.

[4] A. Ban, “Flash file system,” U.S. Patent 5 404 485, April 4, 1995.

[5] A. Ban and R. Hasharon, “Flash file system optimized for page-mode flash technologies,” U.S. Patent 5 937 425, August 10, 1999.

[6] J. Kim, J. M. Kim, S. Noh, S. Min, and Y. Cho, “A space-efficient flash translation layer for compactflash systems,” *IEEE Trans. Consumer Electron.*, vol. 48, no. 2, pp. 366–375, May 2002.

[7] S. W. Lee, D. J. Park, T. S. Chung, D. H. Lee, S. Park, and H. J. Song, "A log buffer-based flash translation layer using fully-associative sector translation," *ACM Trans. Embed. Comput. Syst.*, vol. 6, no. 3, article no. 18, Jul. 2007.

[8] D. Jung, J. U. Kang, H. Jo, J. S. Kim, and J. Lee, “Superblock FTL: a superblock-based flash translation layer with a hybrid address translation scheme,” *ACM Trans. Embed. Comput. Syst.*, vol. 9, no. 4, article no. 40, Mar. 2010.

[9] H. S. Lee, H. S. Yun, and D. H. Lee, "HFTL: hybrid flash translation layer based on hot data identification for flash memory," *IEEE Trans. Consumer Electron.*, vol. 55, no. 4, pp. 2005–2011, Nov. 2009.

[10] M. L. Chiao and D. W. Chang, "ROSE: A novel flash translation layer for NAND flash memory based on hybrid address translation," *IEEE Trans. Computers*, vol. 60, no. 6, pp. 753–766, Jun. 2011.

[11] P. K. Lin, M. L. Chiao, and D. W. Chang, "Improving flash translation layer performance by supporting large superblocks," *IEEE Trans. Consumer Electron.*, vol. 56, no. 2, pp. 642–650, May 2010.

[12] C. Y. Liu, Y. S. Pan, H. H. Chen, Y. C. Wu and D. W. Chang, "Techniques for improving performance of the FAST (fully associative sector translation) flash translation layer," *IEEE Trans. Consumer Electron.*, vol. 57, no. 4, pp. 1740–1748, Nov. 2011.

[13] E. Gal and S. Toledo, "Algorithms and data structures for flash memories," *ACM Comput. Surv.*, vol. 37, no. 2, pp. 138–163, Jun. 2005.

[14] A. Kawaguchi, S. Nishioka, and H. Motoda, "A flash-memory based file system," in *Proc. 1995 USENIX Winter Technical Conf.*, New Orleans, 1995, pp. 155–164.

[15] K. M. J. Lofgren, R. D. Norman, G B. Thelin, and A. Gupta, "Wear leveling techniques for flash EEPROM," U.S. Patent 6 850 443, February 1, 2005.

[16] R. T. Chien, "Cyclic decoding procedure for the Bose-Chaudhuri-Hoc-quenghem codes," *IEEE Trans. Inform. Theory*, vol. 10, no. 4, pp. 357–363, Oct. 1964.

Document analysis based on multidimensional ontology of electronic documents

Viacheslav Lanin

Abstract— The paper describes an approach to semantic indexing indexing of electronic documents based on ontology that describes the structure, type of document and its contents. In addition, existing ontology descriptions of documents are considered and the differences between the proposed multidimensional ontology from them are described. The solution of the problem of analysis of administrative regulations is described as an application of the approach. An algorithm for implementing semantic indexing based on multi-agent paradigm is proposed.

Keywords— multidimensional ontology, semantic indexing, intellectual agents.

I. Introduction

Transition from processing structured data to unstructured data processing is observed in modern information systems. New classes of systems, such as social networking, corporate portals, wiki-resources, etc. became an integral part of the information process. The key point of such systems is "content", which concept can be generalized to "electronic document." Unstructured nature of information raises the question of the transition from traditional indexing documents based on unrelated keywords («bag of words») to the so-called semantic (conceptual) indexing. Semantic (conceptual) document indexing is an indexing, in which synonyms are reduced to the same concept, and disambiguation are separated into different conceptual units [3].

Semantic index of document can become the basis for solving many problems in the processing of electronic documents, in particular, their search, analysis and classification, cataloging and efficient storage, generation and support their life cycle. It's needed to have consolidated knowledge about their structure and content.

Base of semantic index is ontological resource in that following information about the following aspects of electronic documents is needed: electronic document format; type of electronic document; the structure of an electronic document.

When ontological resource is created, it includes concepts related to all three aforementioned aspects of a document information representation. Each of them is described by ontology. Concepts of the various aspects have to be linked. Thus, a single ontology of electronic documents is being created. In addition, the resource should support the ability to expand and specify the settings on the solution of specific problems arising in the processing of documents in a variety of information systems throughout their life cycle.

Thus, in the paper existing ontology resources for describing documents will be examined, an approach to the description of multidimensional ontology will be proposed and an algorithm for semantic indexing based on multi-agent approach will be provided.

II. Related Works

A. Existing Document Ontology

Dublin core [4] – is a set of metadata used to describe documents of various types (publications, audio records, video records). This set specification has status of official international standard (ISO: 15836 2003). The standard has two levels: Simple, comprising 15 elements and Qualified having three additional elements and element refinements (or qualifiers), which refine semantics of the elements. The main feature of Dublin Core is that every element is optional and might be repeated. Dublin Core is a powerful instrument used to describe resources of various types. The fact that it is widespread and flexible is its overwhelming advantage. However, it describes documents tags, i.e. information having indirect correlation with the document content. In this case it is impossible to describe other aspects of the electronic document.

Project ontologies «docOnto» [3] developed by German research group KWARC (Knowledge Adaptation and Reasoning for Content) differ from other projects oriented on formal structure description development (CNXML document ontology) and document semantics (OMDoc document ontology). Members of this group also develop mechanisms of semantic document indexing and tools for document processing. CNXML document ontology (Connexions Markup Language) describes such terms as paragraph, section, reference etc. Ontology is formalized on UML. It gives detailed description of the document. Unfortunately, work in this direction is frozen, last changes date back 2007. One more direction in document ontologies creation is semantics description of documents for narrow subjects, where documents are well formalized, for example mathematical

The reported study was supported by RFBR, research project No. 14-07-31273.

Viacheslav Lanin is with the National Research University Higher School of Economics, Perm, Russia (e-mail: vlanin@live.com).

OMDoc documents. Mathematical Terms, theorems and several other terms are included in ontology.

Document ontology SHOE [5] describes most types of documents. Academic papers are given particular emphasis. Dublin Core reference books and Document Classifier PubMed were the resource.

Document Ontology of Research Centre Linked Data DERI is developed by scholars of Irish Institute DERI (Digital Enterprise Research Institute) and is described in RDFS and OWL-DL [9]. Terms referring project activity documentation are given I the ontology. Developers purposefully refused modelling structure and document content to accommodate flexibility and interoperability.

Muninn project document ontology became the result of processing archive documents of the First World War within the project Muninn WW1 [7]. The Ontology describes bibliography, origins and storage description of the digital item. Most ontology classes are child classes of FOAF. That decision was compatibility possible, on the other hand, make adding additional features of document processing possible, i.e. features for representation document pages, copyright description, etc. One of the main ontology classes is Document, which is integrate class of FOAF Document and Creative Commons Works. Page class describes document pages, in its turn, Image class describes digital page image. Description of different document aspects, document structure in particular, is a significant benefit of this ontology. However, structure description is initially oriented on digital images of archive documents.

Each listed above document ontology has its advantages and disadvantages. We create own ontology specialized on academic paper description.

B. *System for create text-based ontology*

Nowadays there are some information systems that let you create text-based ontology models of documents or let you define correspondence of ontology models thereby transform one model onto another one. We found two web-resources that let you create ontologies: OwlExporter and OntoGrid.

The core idea of OwlExporter is to take the annotations generated by an NLP pipeline and provide for a simple means of establishing a mapping between NLP (Natural Language Processing) and domain annotations on one hand and the concepts and relations of an existing NLP and domain-specific ontology on the other hand. The former can then be automatically exported to the ontology in form of individuals and the latter as data type or object properties [14].

The resulting, populated ontology can then be used within any ontology-enabled tool for further querying, reasoning, visualization, or other processing.

OntoGrid is an instrumental system for automation of creating domain ontology using Grid-technologies and text analysis in natural language.

This system has bilingual linguistic processor for retrieving data from text in natural language. Worth D. derivational dictionary is used as a base for morphological analysis [13]. It contains more than 3.2 million word forms. The index-linking process consists of 200 rules. “Key dictionary” is determined by words allocation analysis in text. The developers came up with new approach of revealing super phrase unities that consist of specific lexical units. The building of semantic net is carried out this way: the text is analyzed using text analysis system, semantic Q-nets are used as formal description of text meaning. The linguistic knowledge base of text analysis system is set of simple and complex word-groups of the domain. This base can be divided into simple-relation-realization base and critical-fragment-set, that let you determine which ontology elements are considered in this text. The next step is to create and develop the ontology in the context of GRID-net. A well-known OWL-standard is used to draw the ontology structure.

III. Using Sample

Consider the example of the proposed approach based on the work with electronic administrative regulations (EAR) [9]. The basic approach to the development of software tools to support the EAR conduct is ontological modeling. Used in the process ontologies are placed in multi-level repository [10], which contains the domain ontology and ontology normative-reference documents. Domain ontology defines the terms used in the documents, namely it describes concepts such as "process", "operation", "artist", etc., in addition, there are included the various classifiers. Ontology of normative-reference documents, in particular, the ontology of the regulation describe the structure of the characteristic elements of documents.

As a result of text description analysis (decomposition) will be built a conceptual model of regulation that, first, to allow it to verify (check structure, identify duplication of information, etc.), and secondly, will link the fragments of a text document with the relevant concepts of the ontology. In addition, the conceptual model of documents could be used to set the "semantic" relationships between different documents and visualization of these links.

Next, consider how ontologies are used in multi-agent semantic indexing algorithm. Domain ontologies used at semantic analysis step. Ontologies that describe the structure of the document (for example, the aforementioned ontology of regulatory-reference documents) are used at the stage of structural analysis. All ontological resources described in RDF-format. Consider in more detail the steps of the analysis of documents used in the algorithm based on semantic indexing agents.

IV. Multi-Agent Semantic Indexing Algorithm

Simplifying the problem we assume that first step of text analysis process was made (for instance using Yandex Mystem[11]), i.e. a set of morphological descriptors for each word have been obtained. All others steps are performed by agent-based semantic indexing. As it could be seen on fig. 1

syntax analysis is not used because it has high time complexity. Instead of this words order in sentence is considered.

Next step is a semantic analysis. The result of the semantic analysis is a semantic descriptor of plain text that binds the morphological descriptors to the elements of the domain ontology. Stop words are skipped.

Next step is a structural analysis. The structural analysis uses document's structure, ontology that describes structure and semantic descriptors of plain text. At this step every concept of structural ontology tries to binds to corresponding structural document element. The result of structural analysis is semantic descriptor of whole text.

Descriptors (morphological, semantic) are a set of tags, which marks each words in the text.

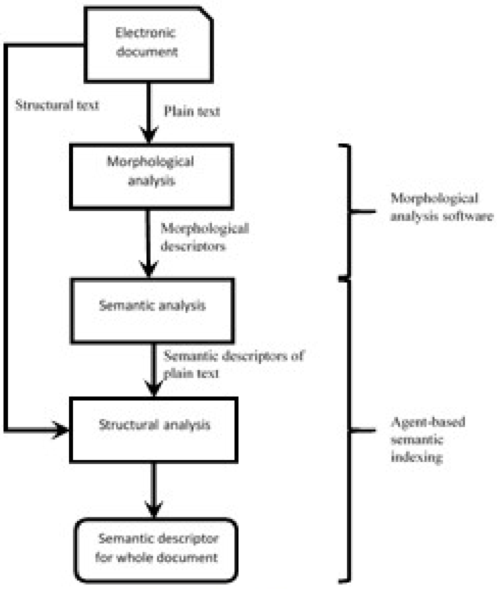

Electronic document
Structural text
Plain text
Morphological analysis
Morphological descriptors
Semantic analysis
Semantic descriptors of plain text
Structural analysis
Semantic descriptor for whole document
Morphological analysis software
Agent-based semantic indexing

Fig. 1. Steps of document analyses

A. *Agent-based solution*

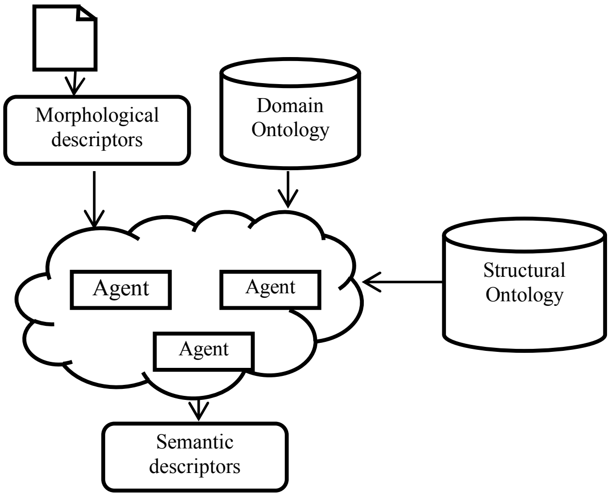


Fig. 2. Steps of solution

Further let us consider the process of building a semantic index based on multi-agent approach (see Fig. 2).

Agents have access to a domain ontology, structural ontology, morphological descriptors and electronic documents which will be indexed. Indexing process is produced on the sentences in the text. Sentences are processed sequentially by agents. The agents form a "team" to index the particular sentence. Thus, agents in the system after the start of the indexing are divided into teams.

B. *Agent Types*

The following types of agents are identified in the system, according to the functional separation:

- Team Lead First Level Agent - TLFL agent,
- Team Lead Second Level Agent - TLSL agent,
- Word Indexer Agent - WI agent,
- Index Writer Agent - IW agent.

The task of WI agent is accessing to the domain ontology and obtaining the set of possible semantic tags for the indexed word. An input word is passed to the WI agent for indexing with the parameters obtained at the stage of morphological analysis. Resulting set of possible semantic tags is passed to the TLSL agent.

TLSL agent binds to morphological descriptors of the sentence and distributes words to all available WI agents. TLSL agent finishes its work on the sentence when the consistent semantic descriptor is formed and written to the document. TLSL agent plans actions for the WI agents and participates in the auction for the resolution of contradictions. After building a consistent semantic descriptor TLSL agent transmits the generated semantic descriptor of the sentence to IW agent who writes semantic tags to the document.

TLFL agent binds to morphological descriptors of the document and distributes descriptors of the sentences to all available TLSL agents. TLFL agent monitors the work of TLSL agents. If the work on the sentence is completed TLSL agent gives TLFL agent a new sentence. In addition, TLFL agent conducts an auction among TLSL agents to resolve ambiguity in the descriptors (see details in section «Agent negotiation»). Besides TLSL agents perform structural analysis. They distribute parts of structural ontology to TLSL agents.

C. *Agent communication*

Agents communicate through language FIPA ACL (Agent Communication Language developed by FIPA) [8]. Two types of actions are used. They are inform (inform about anything) and perform (execution of an action).

Inform action type is implemented in the following cases:

WI agent informs the TLSL agent of completion of indexing word and give it the set of possible semantic tags; content of the communication is as follows: (id, tags), where the id is the identifier word that came to be indexed, tags are returned set of possible semantic tags;

TLSL agent informs the TLFL agent of completion of

indexing sentence with a specific identifier; content of this message contains an identifier of indexed sentence.
Perform action type is implemented in the following cases:

TLFL agent gives to the TLSL agent a task to index a sentence with a specific descriptor; content will look like this: (id, descriptor), where the id is the identifier of the sentence, descriptor is descriptor of the sentence received as a result of syntactic and semantic analysis;

TLSL agent gives a task to the WI agent to index a word with specific id; content will look like this: (id, word, parameters), where id is ID of the word, word is the word for indexing, parameters are parameters obtained at the stage of morphological and syntactic analysis;
TLSL agent gives a task to the IW agent to write semantic tag of specific word; content is as follows: (word, tag), where the word is an indexed word, tag is just a semantic tag of indexed word.

D. Planning

The planning is dynamic. TLSL agents themselves form a team of agents from the available WI agents. A count of needed WI agents depends on structure of a sentence. With a lack of WI agents at the time of formation of the team TLSL agent may designate to perform indexing of few words at once to the same WI agent. TLFL agent monitors the performance of work of TLSL agents and if they are released it assigns them new sentences for indexing. Completing of work of the agents (WI and TLSL) monitored not only by sending their corresponding messages of inform type, but also change their states (agent states) in the meaning of "vacant."

E. Agent knowledge bases

WI agents and IW agents are primitive reflex agents working in the mode of stimulus-response. Their main function is a simple, no inference, execution of work. In the knowledge bases of these agents are only procedural steps.

Knowledge bases of TLFL and TLSL agents represent productions with embedded procedural actions. In fact, the script actions are necessary for the distribution of work between agents. Accordingly TLSL agent knowledge base contains a script for word distribution among WI agents, and TLFL agent knowledge base includes a script for sentences distribution between agents TLSL.

F. Agent negotiation

TLFL agent conducts an auction among agents TLSL, each of which has a contextual memory (training component). Every TLSL agent using the contextual memory votes for a one option of sematic descriptor of the sentence. Option of semantic descriptor of the sentence with the highest number of votes will be considered as a true semantic descriptor of the sentence. The set of all consistent semantic descriptors of the sentences form the document semantic descriptor.

V. Conclusion

Unlike existing ontologies describing documents multidimensional ontology represents the document structure, which allows to consider this information during indexing process. In developing ontologies it included the mechanisms for integration with domain ontologies and expanding of ontology - adding new "aspects", which also expands the scope of the decision. The proposed multiagent approach creates preconditions for solving the optimization problem of parallel execution of semantic indexing.

Also planned that the developed ontology and algorithms will be used in a number of projects related to the development of domain-specific languages (Domain Specific Languages, DSL) for different domains based on linguistic tools MetaLanguage.

References

[1] Segaran T., Evans C., Taylor J. Programming the Semantic Web, O'Reilly Media, 2009.

[2] Lukashevich N.V., Dobrov B.V. Bilingual information retrieval based on the automatic conceptual indexing // Computational linguistics and intelligent technologies. Proceedings of the International Conference "Dialogue-2003". Protvino. June 11-16 2003y. / Ed. by I.M.Kobozevoy, N.I.Laufer, V.P.Selegeya - M.: Science, 2003. - pp.425-432.

[3] CNXML/DocumentOntology http://mathweb.org/wiki/CNXML/DocumentOntology

[4] Dublin Core Metadata Element Set, Version 1.1 http://dublincore.org/documents/dces/

[5] Document Ontology (draft) http://www.cs.umd.edu/projects/plus/SHOE/ onts/docmnt1.0.html

[6] Grishman. R. TIPSTER Architecture Design Document Version 2.3. Technical report, DARPA, 1997. http://www.itl.nist.gov/div894/894.02/related_projects/tipster/.

[7] Muninn Documents Ontology http://rdf.muninn-project.org/ontologies/documents.html

[8] XML Languages http://cnx.org/help/authoring/xml

[9] Lanin VV Lyadova LN Technology of support maintenance of electronic administrative regulations based on ontological models / / Proceedings of the All-Russian conference with international participation "Knowledge-Ontology-Theory." Novosibirsk, 2011, pp. 38-46 V.2.

[10] Lanin V.V. Using multi-level ontology repository for electronic document analysis / / Proceedings of international scientific conference "Intelligent systems» (AIS'08) and "Intelligent CAD» (CAD-2008). Scientific publication in 4 vols. T. 1. - Moscow: Fizmatlit 2008. Pp. 202-206.

[11] Program for morphological analysis of text in Russian "Mystem". [Electronic resource] [Mode of access:http://company.yandex.ru/technologies/mystem/] [Checked at: 24.06.12]

[12] D. Worth, A. Kozak, D. Johnson "Russian Derivational Dictionary", New York, NY: American Elsevier Publishing Company Inc, 1970

[13] R. Witte, N. Khamis, and J. Rilling, "Flexible Ontology Population from Text: The OwlExporter" Dept. of Comp. Science and Software Eng. Concordia University, Montreal, Canada. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/932_Paper.pdf

The Target vs. Non-Target Classification Approach for Biometric Recognition Applications

Sorin R. Soviany, Sorin Puşcoci and Cristina Soviany

***Abstract*—**The paper approaches the biometric identification with a special kind of classifiers named detectors. This approach relies on a target vs. non-target biometric data classification strategy; this is suitable for applications in which there are only a few most authorized end-users and for which their accurate identification is very important. Therefore the detectors-based biometric data classification provides a reliable solution for biometric recognition systems design with several security levels. The performance improvement for identification is supported by the detectors design principle with a target class-based threshold. Although not all the biometric applications could benefit from this approach, there are obvious advantages in design biometric data classifiers for a few target identities with low complexity classification models running on reduced feature spaces.

***Keywords*—**detectors, identification, target class.

I. Introduction

THE biometric identification still remains an important challenge for various security applications, despite of the significant advances either in biometric sensors technologies but also in pre-processing and processing algorithms for individuals recognition. This is because of the huge space of the possible identities for the large-scale identification systems. An identification system has to guess who a real individual is only based on his/her biometric credential, without any additional identifiers like usernames or other IDs. The false acceptance/rejection error rates are still larger in identification mode than in verification mode. Also there are applications with several security levels and for which it is important only to accurately identify a few persons (the most authorized users of a medical database, for instance). In these cases the optimal trade-offs between identification accuracy and computational complexity should be found to provide an efficient solution.

The related works so far performed in biometric systems design concern the 2 main stages in biometric data processing: feature generation (providing the biometric templates) and data classification/matching (providing acceptance/rejection or identification decisions, respectively). For both of them there are a lot of developments supporting the actual commercially available biometric systems. Regarding to the classification stage, many researchers focused on the 2 main approaches, either for unimodal or multimodal biometrics: distance-based classification and learning-based classification. The distance-based classification (or matching) approach is commonly applied in many already developed systems, and typically the authors used various matching scores normalization techniques, such as those given by Ross and Jain in [1]. The most applied distances for matching score computation are Euclidian and Mahalanobis [2]. However the matching process is still computational expensive because of the typical huge searching space for the large-scale identification systems. Although additional techniques were applied in order to optimize searching within the possible identities space [3], there are still challenging issues for identification in applications with several security levels and for which the complexity remains a challenge. As concerning the learning-based classification, there are a lot of achievements with various classifiers which are applied for biometric datasets with different feature space sizes, with or without additional feature spaces transformations such as PCA (Principal Component Analysis) or LDA (Linear Discriminant Analysis) [4]-[6] On the other hand, the multi-class nature of the biometric identification task was intensively approached with various classification models. Neural networks were applied in biometric recognition, especially for identification task [7]. SVM classifier was applied in biometric applications, for one-class or two-class problems [8].

Most of these achievements improved the accuracy for verification as much as this biometric process is typically approached with a one or two-class classification strategy. The identification process could be however approached in a similar way if the application requires several authorization degrees for the end-users and in this case the design solution is based on detectors, which are only trained for target vs. non-target identification decisions. Therefore we design a classification system for biometric data in which the training is only performed on some target identities. This classification

Sorin R. Soviany, Ph.D. in Electronic and Telecommunications, is a senior research engineer within the National Communications Research Institute (I.N.S.C.C.), Bucharest, Romania (office phone: +40-21-3000011 e-mail: sorin.soviany@ inscc.ro).

Sorin Puşcoci. Ph.D. in Electronic and Telecommunications, is Head of Communications Terminals and Telematic Department and senior research engineer within the National Communications Research Institute (I.N.S.C.C.), Bucharest, Romania (e-mail: sorp2006@gmail.com).

Cristina Soviany is Founder, CEO for Features Analytics SA, Bruxelles, Belgium (e-mail: cristina.soviany@ides-technologies.com).

model could be easily integrated within a hierarchical classifier for a multi-level security system; in this case, each identification decision stage defines a security level.

The remainder of this paper is structured as follows. Section II presents the essentials of the proposed method for target vs. non-target classification of biometric data, focusing on the detectors design principle. In section III the achieved experimental results are presented. Section IV concludes our paper.

II. THE TARGET VS. NON-TARGET CLASSIFICATION METHOD FOR BIOMETRIC RECOGNITION

A. The System Architecture

We apply the target vs. non-target classification method for a palmprint-based biometric system with the functional architecture given in fig 1.

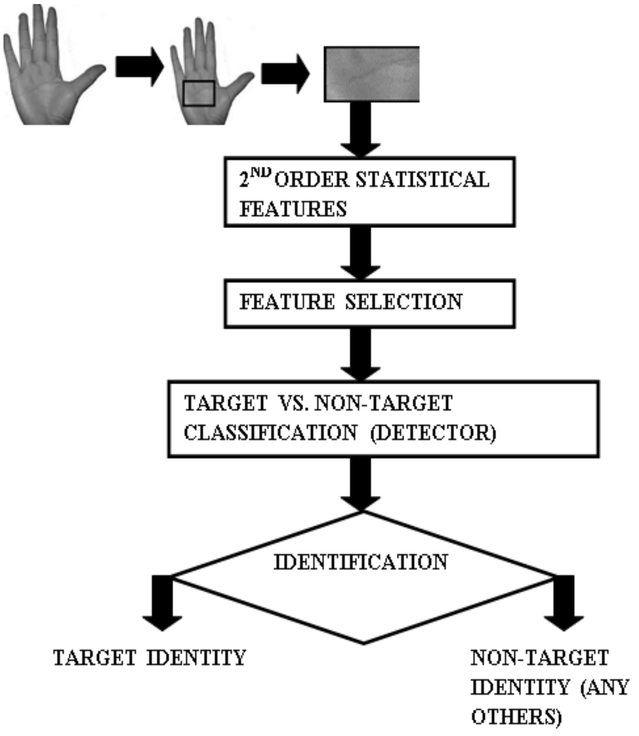

2ND ORDER STATISTICAL FEATURES
FEATURE SELECTION
TARGET VS. NON-TARGET CLASSIFICATION (DETECTOR)
IDENTIFICATION
TARGET IDENTITY
NON-TARGET IDENTITY (ANY OTHERS)

Fig. 1 The functional architecture for the palmprint-based identification system

Our available palmprint dataset is provided from 50 individuals from which we selected the most important user to be recognized with a high priority using a detector-based classification method (target vs. not-target). The main focus of our research is on the classification stage providing identification decisions.

B. Feature Generation : 2^{nd} order statistical feature extraction and selection

For this step we applied a **regional and textural approach based on co-occurrence matrices** to derive the required features. The 2^{nd} order statistical features provide information regarding the statistics of the gray-level distribution among pixels pairs, with 2^{nd} order histograms or co-occurrence matrices [9]-[11]. The textural approach with co-occurrence matrices was applied for face recognition systems [10]; in our experiments we generate these statistical features for the palmprint. The feature extraction is done on a previously selected region of interest (ROI) within the original palmprint image. Other additional statistical features which are derived from the applied co-occurrence matrix are the following 4 measures: angular second moment, contrast, inverse difference moment and entropy for pixels pairs [9]. The resulted features number (feature vector size) L_X for our experimental palmprint dataset is given by

$$L_X = size(X) = \left(GLB\right)^2 + 4 \qquad (1)$$

where:
X is the resulted feature vector for the palmprint image;
GLB is the gray-level bins for which we computed the co-occurrence matrix. For our experiments, the value of this parameter is fixed to $GLB = 4$, providing an initial feature space dimensionality of *20* (*16* plus the *4* additional statistical features). The reasons for this relatively small value of GLB (which is the 1st parameter of our feature extractor) are the following:

- it provides a significant feature space dimensionality reduction, allowing to apply low-complexity classifiers for the biometric samples. According to Theodoridis [9], a low-dimensional feature space is more efficiently covered with a small number of training samples. This is important for many biometric applications in which the available training datasets for system design are reduced in order to maintain the end-users acceptability;
- it provides a co-occurrence matrix with less null elements, therefore the generated feature sets will mostly contain significant values for the biometric patterns separation in order to support an accurate identification task.

However, this information loss could be balanced with a lower displacement distance value (the pixels number between the pixel pairs in co-occurrence matrix evaluation); this is the 2nd parameter of this feature extractor. For our data this fixed value is 2. This parameter should not exceed a certain threshold (in our case, this threshold value was 4), otherwise the overall pixels pairs number will reduce and therefore we would have a small amount of useful information for the biometric identification.

We also applied a **feature selection** algorithm for a further feature space dimensionality reduction. We evaluated several non-optimal and non-exhaustive feature selection algorithms on our experimental dataset: *forward-* and *backward-searching, floating-search, random searching and individual* *ranking feature selection*. For features evaluation we applied a wrapper approach in which the performance criterion was *1-NN classification error rate minimization*. The main reason for this choice is that the 1-NN classification rule is asymptotically at most 2 times as bad as the Bayes decision rule [12],[13]. Finally we selected the *individual ranking* algorithm because of its high speed on our available palmprint data; its execution time was almost 3 times lower than for the sequential searching algorithms and almost 2 times lower than for the random searching feature selection. The resulted optimal palmprint feature vector size was *11*. Therefore we applied the target vs. non-target classification method for a palmprint feature space with *11* dimensions.

C. Biometric Data Classification: Target vs. Non Target Identification with Detectors

The palmprint samples are further processed with the proposed classification method in which the classifiers outputs are focused on a target identity. This classifier is called detector [12]; the model is built typically through its output thresholding for a certain target class (enrolled identity). Actually a biometric detector training is performed in the following way:

- all the training samples belonging to the target enrolled identity are grouped in one class;
- all the other training biometric samples belonging to all the other non-target identities are grouped in the 2nd class;
- training the detector to only identify the target person, which is the most authorized end-user of the protected resource.

The differences between the detector-based classification (target vs. non-target) and discriminant-based (multi-class) classification are depicted in figures 2 and 3 for the (simplified) 2-dimensional case, respectively. In fig. 2 one can see that there is a focused decision region for one target identity, while the 2nd decision region contains the biometric datapoints from all the other non-target identities. In fig. 3, for the 3 identities there are 3 decision regions, respectively.

Therefore the multi-class problem of the biometric identification could be easily approached with a 2-class classifier, but still performing identification. This identification is focused on a certain individual and this is the main difference between the detectors-based approach and the typical models which are based on the multi-class approach; in the last case, each class contains all the biometric samples provided by a certain enrolled individual.

For the available datasets we used Parzen models with Laplace kernel. The estimation model (class-conditional probability density) is given by [9], [12], [14]

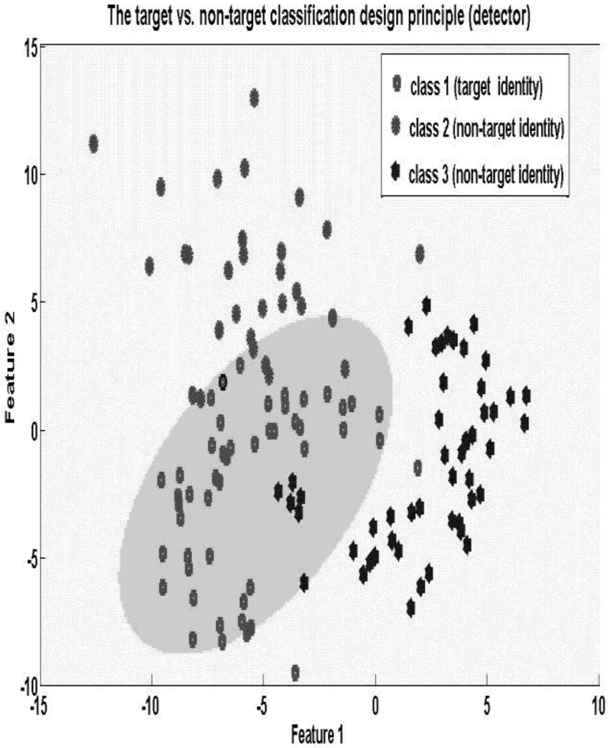


Fig. 2 The target vs. non-target classification design principle (the detector classifier for focused identification)

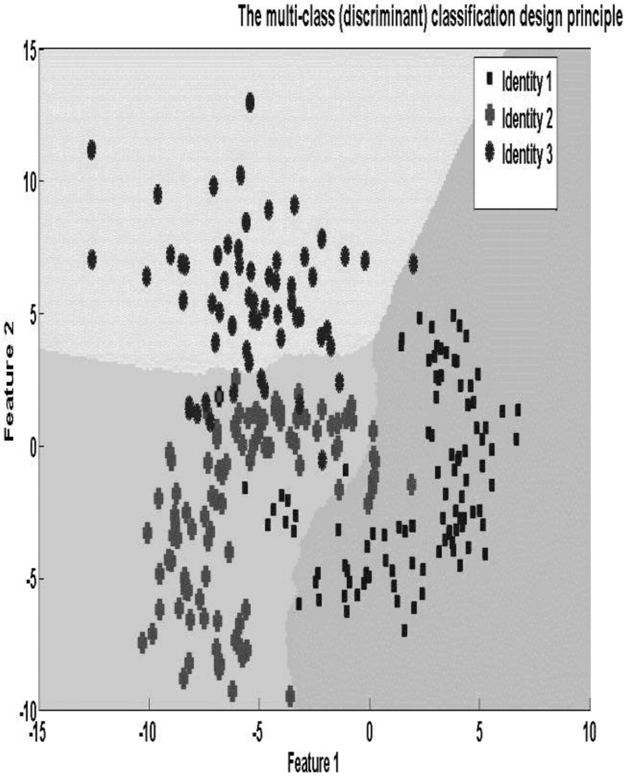


Fig. 3 The multi-class classification design principle (the discriminant classifier for 3 enrolled identities)

$$P(X \mid I_X) = \frac{1}{N_{I_X}} \cdot \sum_{Z_{I_X} \in I_X} K_h\left(\frac{X - Z_{I_X}}{h}\right) \qquad (2)$$

where:
X is the current datapoint in the feature space (the biometric sample provided by an individual for his/her identification)
I_X is the class label. For a **detector** model ***(target vs. non-target***), the 2 classes are the following:

$$I_X = \begin{cases} I_T, \text{ for the target identity class label} \\ I_{non-T}, \text{ for all the non-target identities class label} \end{cases} \qquad (3)$$

For a **discriminant** model (multi-class biometric identification for *M* enrolled individuals), the class labeling decisions are:

$$I_X = \begin{cases} I_1, \text{ for the } 1^{st} \text{ enrolled identity} \\ I_2, \text{ for the } 2^{nd} \text{ enrolled identity} \\ \ldots \\ I_M, \text{ for the } M^{th} \text{ enrolled identity} \\ I_?, \text{ for not enrolled individual (unknown)} \end{cases} \qquad (4)$$

N_{I_X} is the number of biometric samples belonging to identity I_X;
Z_{I_X} is the training feature vector drawn from the class I_X distribution;
h is the smoothing parameter for the kernel function K_h. From the available datasets we found on optimal value of 0.65;
K_h is the kernel (multi-dimensional) function centered around the component Z_{I_X} of the training dataset. We used ***Laplace kernel*** because it is less sensitive to changes in variances/co-variances than the ***Gaussian kernel*** [12].

In this case, the Bayesian decision function for the target vs. non-target classification with Parzen based model and Laplace kernel becomes:

$$g(X) = P(I_T) \cdot P(X \mid I_T) - P(I_{non-T}) \cdot P(X \mid I_{non-T}) \qquad (5)$$

in which the prior probability estimators are evaluated as follows:

$$P(I_X) = \frac{N_{I_X}}{N}, I_X \in \{I_T, I_{non-T}\} \qquad (6)$$

N is the number of all biometric samples, no matter their class (identity) membership. Therefore the detector model for the target vs. non-target identification is:

$$g(X) = \frac{1}{N} \cdot \left[\sum_{Z_{I_T} \in I_T} K_h\left(\frac{X - Z_{I_T}}{h}\right) - \sum_{Z_{I_{non-T}} \in I_{non-T}} K_h\left(\frac{X - Z_{I_{non-T}}}{h}\right) \right] \qquad (7)$$

Finally, as much as the detection (target vs. non-target classification) is a thresholding-based approach, the decision rule for target vs. non-target identification is

$$ID(X) = \begin{cases} I_T, \text{ if } g(X) \geq \theta_T \\ non - I_T, \text{ otherwise} \end{cases} \qquad (8)$$

where θ_T is the target identity threshold.

D. The Hierarchical Classifier

The detector classification model for biometric data (target vs. non-target identification) could be easily integrated within a hierarchical classifier with several decision levels, in which each decision level defines a certain security layer for the application. Fig. 4 depicts an example with 2 decision (security) levels.

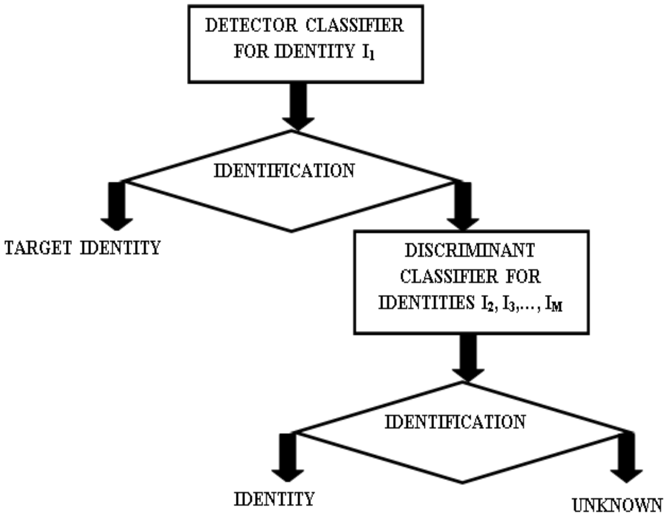

DETECTOR CLASSIFIER FOR IDENTITY I1
IDENTIFICATION
TARGET IDENTITY
DISCRIMINANT CLASSIFIER FOR IDENTITIES I2, I3,..., IM
IDENTIFICATION
IDENTITY
UNKNOWN

Fig. 4 The multi-class classification design principle (the discriminant classifier for 3 enrolled identities)

We could define various hierarchical classifiers, either homogeneous (with the same underlying model) or heterogeneous (with various models applied for biometric data recognition). A critical issue is the execution time, which relates to the hierarchy depth. This is especially important for large-scale identification systems and it is a reason for working with reduced feature space dimensionalities while applying such classification models for biometric identification applications. Further research should be done while considering this timing issue.

III. EXPERIMENTAL RESULTS

We evaluated this classification method for palmprint data with a feature space having only 11 dimensions. First we trained the classifiers for several values of the smoothing parameter, finally resulting in an optimal value of 0.65. Actually we manually selected the following values of this parameter: 0.45; 0.50; 0.55; 0.60; 0.65; 0.70. The performance measure was average identification error rate on 15 experiments. The learning curves are depicted in fig. 5 and they allowed also to set the training set size such as to achieve a suitable classifier behavior on the available data. We applied the same number of biometric samples per class (target and non-target, respectively). The class unbalancing problem for biometric recognition will remain to be approached in a future paper.

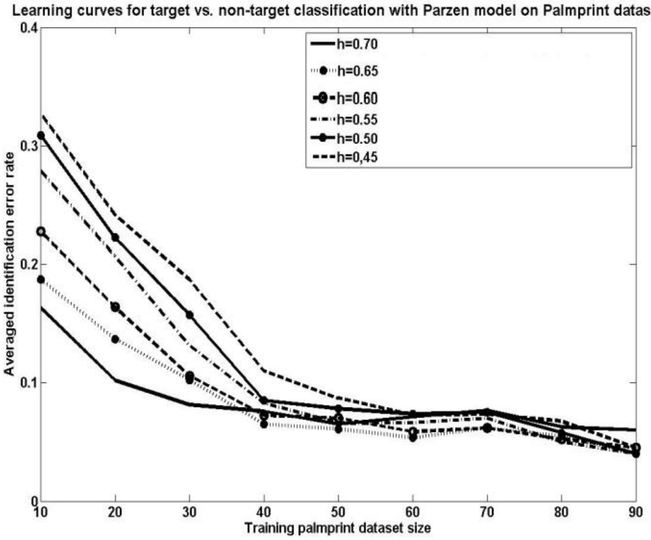


Fig. 5 Learning curves for various smoothing parameters

For training the classifiers with a number of samples per class not exceeding 50, the best models are achieved for h = 0.65 and for h = 0.70; for 40 samples per class, the optimal value resulted to be h = 0.65.

On the other hand, in this process we applied a leave-one-out cross-validation in order to enhance the generalization capacity of the proposed solution for palmprint data classification.

The next step is to find out the optimal operating point for the detector classifier (target vs. non-target) on the experimental palmprint dataset, for some rejection thresholds. A classification rejection threshold is very useful in biometric applications because typically not all the individuals are always able to provide suitable biometric samples. Therefore it is important to reject those biometric samples with a high probability to generate identification error rates. This approach allows to further optimize the identification process while minimizing the errors caused by low-quality input biometric data. Actually the optimization looks to find out the best trade-off between the performance and the security threshold for the given application.

The TPR (True Positive Rate) vs. rejection thresholds curve is represented in fig. 6.

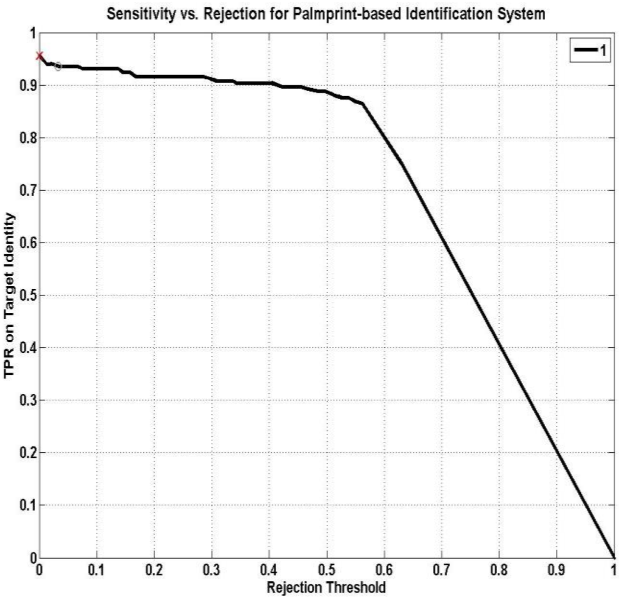


Fig. 6 The system sensitivity vs. classification rejection thresholding

The typical rejection threshold in biometric applications is around 5%. This means that almost 5% from all the test users are not able to provide the high quality biometric samples. In our experiments, for a rejection fraction of 4%, the achieved TPR on the target class (the focused identity to be recognized) is 0.94. This measure is evaluated just on the target class, therefore on the most important individual that should be precisely identified. On the other hand this result is achieved for a feature space with a smaller dimensionality than other actual developments on palmprint-based biometric systems, either for verification or for identification working modes.

IV. CONCLUSIONS

A target vs. non-target classification approach for individuals biometric recognition should be a reliable solution when the applications require several security levels. The main difference from most of the actual data classification approaches for biometrics (either unimodal or multimodal) is just this focus on a target identity to be recognized. This could be a very convenient method especially when not all the enrolled individuals should be identified with the same accuracy.

On the other hand, in our experiments, either for a single detector-based classification system or for hierarchical classifiers with several detection and discrimination stages, we worked with small-dimensional feature space, in this case for palmprint data. A suitable reduced feature space should allow to apply low-complexity classifiers for biometric data, but with a carefully feature selection procedure in order to maintain the most discriminative and informative features for individuals recognition.

One of the issues that should be considered when the security application requires to design a biometric system with a hierarchical classifier is the execution time, which is related to the hierarchy depth, and its impact on the overall identification accuracy. So far we evaluated several hierarchical classifiers with only 2 and, respectively, 3 decisions levels, but without an estimate of the execution time.

Another issue is concerning the feature sets. Here we applied a simplified procedure for feature extraction with 2nd order statistical features based on co-occurrence matrices. The main reason for our approach was that it allowed us to easily adjust the features number in order to provide a significant dimensionality reduction but maintaining the identification accuracy for the target identity.

Finally, we should remark that the achieved results are strongly depending on the available data. It is not possible to design the same best solution for all the applications. The various security applications have not the same requirements and implementation costs, and therefore the designed system should be carefully optimized to the application.

.

References

[1] Jain A., Nandakumar K., Ross A.: Score Normalization in multimodal biometric systems, Pattern Recognition, The Journal of the Pattern Recognition Society, 38 (2005).

[2] Supriya Kapoor, Shruti Khanna, Rahul Bhatia: Facial Gesture Recognition using Correlation and Mahalanobis Distance, *(IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, Nr.. 2, 2010*

[3] Mhatre Amit, Palla Srinivas, Chikkerur Sharat, Govindaraju Venu: Efficient Search and Retrieval in Biometric Databases, Biometric Technology for Human Identification II. Edited by Jain, Anil K.; Ratha, Nalini K. Proceedings of the SPIE, Volume 5779, pp. 265-273 (2005)..

[4] Yagiz Sutcu, Shantanu Rane, Jonathan Yedidia, Stark Draper, Anthony Vetro: Feature Transformation of Biometric Templates for Secure Biometric Systems Based on Error Correcting Codes, *MITSUBISHI ELECTRIC RESEARCH LABORATORIES , 2008*.

[5] Hyunsoo Kima, Barry L. Drake, Haesun Park: Multiclass Classifiers Based on Dimension Reduction with Generalized LDA, 2006

[6] Tao Li, Shenghuo Zhu, Mitsunori Ogihara: Using Discriminant Analysis for Multi-class Classification: An Experimental Investigation, 2009

[7] M. Gopikrishnan, T. Santhanam. Effect of different Neural Networks on the Accuracy in Iris Patterns Recognition, International Journal of Reviews in Computing 30th September 2011. Vol. 7, 2011

[8] D.Kumar, P.Unikrishnan. “Class Specific Feature Selection for Identity Validation using Dynamic Signatures”, J.Biomet Biostat, JBMBS, Vol. 4, Issue 2, 2013

[9] S.Theodoridis, K Koutroumbas., “Pattern Recognition 4th edition” Academic Press Elsevier, 2009.

[10] A. Eleyan. H. Demirel.: “Co-occurrence matrix and its statistical features as a new approach for face recognition”, Turk J Elec Eng & Comp Sci, Vol.19, Nr.1, 2011.

[11] Bino S. V, A. Unnikrishnan and Kannan B.: “Gray level Co-Occurrence Matrices: Generalisation and some new features”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[12] .*** PerClass Training Course: Machine Learning for R&D Specialists, Delft, Netherlands.

[13] Devroye L., Gyorfy L., Lugosi G.: “A Probabilistic Theory of Pattern Recognition”, Springer, 1997.

[14] D. Zhang, F.Song Y.Xu, Z.Liang, “Advanced Pattern Recognition Technologies with Applications to Biometrics” , Medical Information Science Reference, IGI Global, 2009

Sorin R. Soviany was born in Bucharest, Romania, in 1976. He graduated the Politehnica University of Bucharest, Faculty of Computer Sciences and Control Engineering in 2000 (M.Sc.). Since 2013 he is a Ph.D. in Electronic and Telecommunications Engineering at University of Pitesti, Romania, with the thesis *Decision Optimization in Biometric Identification Systems*.

Actually he is a senior research III engineer at National Communications Research Institute (I.N.S.C.C.), Bucharest, Romania. He is involved in research project focused on data and networks security and with applications for telemedicine systems. Some relevant published papers are the following:

Sorin Soviany, Sorin Puşcoci: *An Optimized Multimodal Biometric System with Hierarchical Classifiers and Reduced Features*, IEEE International Symposium on Medical Measurements and Applications MeMeA, Lisbon, 11-12 June, 2014;

Sorin Soviany, Cristina Soviany: *A Biometric Security Model with Identities Detection and Local Feature-level Fusion*, The 2013 International Conference on Security and Management (SAM'13), World Academy of Science, SUA, Las Vegas, 22-25 July, 2013

Sorin Soviany, Cristina Soviany, Mariana Jurian: *A Multimodal Approach for Biometric Authentication with Multiple Classifiers*, International Conference on Communications, Information and Network Security (ICCINS 2011), Venice, Italy, 28-30 november, 2011

His major research topics include networks security technologies, applications of modern pattern recognition algorithms especially for biometric systems, communications networks reliability, image processing for biometric data.

Dr. Sorin R. Soviany is member of Romanian Engineers General Association (A.G.I.R.) and Romanian Medical Informatics Society.

Sorin Puşcoci was born in Bucharest, Romania, in 1953. He is Diplomat engineer, electronic and telecommunications, Polytechnic Institute Bucharest – Faculty of Electronic and Telecommunications, 1977. Since 2007 he is Ph.D. in Electronic and Telecommunications Engineering at University of Pitesti, Romania, with the thesis *Contributions about implementations of telemedicine national network*.

Actually he is senior research II, at National Communications Research Institute- I.N.S.C.C., head of Communications Terminals and Telematic Department. He managed and is involved in research projects focused on telemedicine development and applications. Some relevant papers:

V. Stoicu-Tivadar, L. Stoicu-Tivadar, **S. Puşcoci**, D. Berian,V. Topac: *Webservice-based solution for an intelligent telecare system*, Individual Book Chapter in Studies in Computational Intelligence, Volume 378/2012, 383-408, DOI: 10.1007/978-3-642-23229-9_18;

S. Puşcoci, L. Stoicu-Tivadar, V. Stoicu-Tivadar, D. Berian, F. Serbanescu, S. Ionita, F. Bajan: *Integrated teleassistance platform with enhanced accessibility to information* – TELEASIS, 6th IEEE International Symposium On Applied Computational Intelligence And Informatics (SACI 2011), May 19-21, 2011 In Timisora, Romania

Cristina Soviany was born in Bucharest, Romania, in 1968. She graduated the Politehnica University of Bucharest, Faculty of Computer Sciences and Control Engineering in 1991 (M.Sc.).Since 2003 she is Ph.D. in Applied Sciences, Delft University of Technology, Faculty of Applied Physics, Pattern Recognition Group, Netherland, with her thesis *Embedding Data and Task parallelism in image processing applications*.

Founder and CEO of IDES Technologies (December 2009), based in Braine-l’Alleud, Belgium; **Since May 2014** - Founder, CEO for Features Analytics SA. Some relevant papers:

S.K.Govindaraju, Mark Emberton, Hashim Ahmed, Mahua Sahu, **Cristina Soviany**, Harry Bleiberg, Rina Nir, Dror Nir, “ Prostate HistoScanning’s role in visualizing early disease”, in *European Urology Today*, vol.21 – no. 1 February/March 2009

Her main research topics and interest are image processings, advanced pattern recognitions algorithms for medical and security applications.

She has received an award for the most innovative Information and Communication Technology European startup company in December 2011

Active member of international scientific conferences program committees.

2014: Member of Program Committee at The 2014 International Conference on Security and Management, Las Vegas, USA

A Survey on Mobile Augmented Reality Based Interactive Storytelling

Sagaya Aurelia, Dr. M. Durai Raj, Omer Saleh

Abstract**--**Mobile technology improvements in built-in camera, sensors, computational resources and power of cloud sourced information have made AR possible on mobile devices. This paper surveys the field of mobile augmented reality and how it is used as interactive, collaborative and location based story telling medium. This survey provides a starting point for anyone interested in researching or using Mobile Augmented Reality and interactive storytelling irrespective of the application.

Keywords:Human Computer Interaction,Immersive environment, Interactive storytelling,Mobile Augmented reality.

I. INTRODUCTION

This paper surveys the current state-of-the-art in Mobile Augmented Reality and how it is used as a medium for Interactive, Collaborative and location based storytelling.

A survey paper does not present new research results. The contribution comes from consolidating existing information from many sources and publishing an extensive bibliography of papers in this field. This paper provides a good starting point for anyone interested in beginning research in this area[2].

Section 1 describes what MAR is, and Interactive storytelling. Section 2 explains the related works based on interactive book. Multimodal, Multi-User and Adaptive Interaction for Interactive Storytelling, Location based storytelling, Storytelling in Collaborative Augmented Reality Environments. Finally, Section 3 draws reviews and conclusions, and describing some applications where the combination of Mobile Augmented reality and Interactive storytelling medium can be used.

II. DEFINITION

1. *Augmented Reality:*

Extend Azuma's [Azuma et al. 2001] definition of AR to MAR in a more general way as follows:

- Combine real and virtual objects in a real environment.
- Interactive in real time.
- Register (align) real and virtual objects with each other.
- Run and/or display on a mobile device.

We do not restrict any specified technology to implement a MAR system. Any system with all above characteristics can be regarded as a MAR system.

We do not restrict any specified technology to implement a MAR system. Any system with all above characteristics can be regarded as a MAR system. A successful MAR system should enable users to focus on application rather than its implementation [Papagiannakis et al. 2008]. MAR is the special implementation of AR on mobile devices. Due to specific mobile platform requirements, MAR suffers from additionalproblems such as computational power and energy limitations. It is usually required to be self-contained so as to work in unknown environments. AR and MAR supplement real [1].

III. INTERACTIVE STORYTELLING

Interactive storytelling is a form of digital entertainment where authors, public, and virtual agents participate in a collaborative experience. Crawford [2004] defines interactive storytelling as a form of interactive entertainment in which the player plays the role of the protagonist in a dramatically rich environment. The experience offered to the public by an interactive story differs substantially from a linear story. An interactive story offers a universe of dramatic possibilities to the spectator. In this form of entertainment, the audience can explore an entire set of storylines, make their own decisions, and change the course of the narrative.

Typically, the way viewers interact with a storytelling system is directly linked to the story generation model: character or plot-based model. Character-based approaches [Cavazza et al. 2002][Young 2001][Aylett et al. 2006] give to the system great freedom of interaction. Usually, the story is generated based on the interactions between the viewer and the virtual characters. In some cases, the viewer can acts as an active character in the story. In plot-based approaches [Grasbon and Braun 2001][Ciarlini et al. 2005], the interaction options are quite limited. The users can perform only subtle interferences to guide the progress of the narrative plot[3].

1. *Interactive Augmented reality storybook:*

Listening to stories draws attention to the sounds of language and helps children develop a sensitivity to the way language works'', said by Isbell [14]. She stated that children tend to learn more while they are listening to stories because stories are fiction which most of it are fairytales that will not happened in the reality. It challenged children's imagination, hearing and seeing while listening to stories. Therefore, AR was brought in to build a system which include audio and graphics which allows children not only to read aloud with but interact with the system at the same time [4].

1.1 Multimedia Interactive Book (miBook)

Compared to previous AR book setup which needed some initial setup before usage, this mobile AR book concept can be used directly as a normal user reading a normal book. Thus, it will reduce the learning hassle as well

Mrs P. Sagaya Aurelia is with Barathidasan university as part time research scholar, presently working in Azzaytuna university, Libya (e-mail: psagaya.aurelia@gmail.com).

Dr. Durai Raj is with Department of Computer Science, Bharathidasan university, Trichrapalli, India(phone no: 0919487542202,email:durairaj.bdu@gmail.com)

Dr Omer Saleh is currently the Director of Post graduate, research and Development cum Head of the Department, Department of computer science, Faculty of Education, Azzaytuna University, Libya(phone no: 0926895760,e-mail:Immer.jomah@gmail.com)

as increasing the interactivity between user and book. This so called playbook is mainly focused on two parts which consist of physical book and mobile application. Mini Interactive Book is a new tool providing a responsive environment and an interactive learning, which handles with different type of content.

miBook is the combination of a printed book (or its digital format) with the respective audiobook and its 3D models (as well as the 2D graphics). Technologically, miBook environment consists of a handheld camera, a personal computer (to generate user's individual scene views), and a physical book. miBooks uses "normal books" with text and pictures on each page and have an additional audio content – the correspondent audiobook.

By supporting a real-time AR texture-tracking algorithm, which uses the novel feature detection technique from Bastos and Dias (2007) (see Figure 1), the enhancement of global algorithm performance allows the support of different hardware profiles, both in desktop and mobile setups. It also includes the possibility of tracking several images/textures at the same time and it supports several 3D standard formats (3DS, VRML, OBJ, DXF, Cal3D, among others). As we can see on Figure 1, there is no need to have the black borders as tracking marks. The first picture on Figure 1 (left side) is the 2D sheet of a book and the right side one shows a 3D object registration where someone is interacting in real time with miBook. As for interactivity enhancement, miBook features provide a physic engine to enable scientific simulations. It will also enable audio storytelling with virtual elements interaction (Script) and both artificial intelligence and speech recognition algorithms for user guidance. All features may be available both in desktop and mobile (PDA or Smartphones) setups, being one of the biggest breakthroughs for the AR community [5].

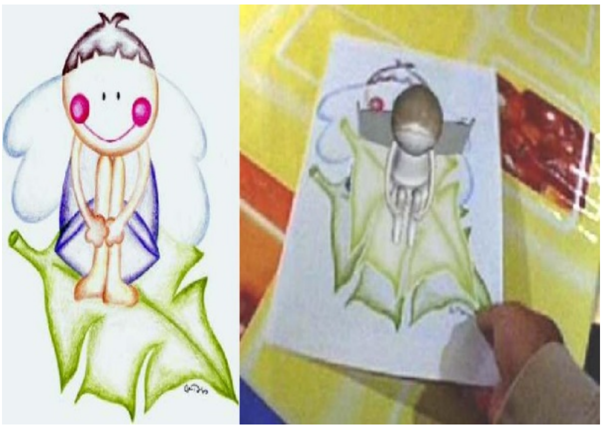


Fig. 1 Example of virtual object registration in a real scene in miBook9(texture image and registered scene)[5]

The application of the miBook solution to new forms of learning can be naturally and fully under control of users (both students and educators). The new interactive way of linking traditional pedagogical approaches (such as reading printed books), common devices capabilities (like handheld devices with camera) and the potential of multimedia technologies (audiovisual interpretation technologies) can provide a better understanding, knowledge acquisition and enhanced learning experience[5].

1.2 An Interactive Mobile Augmented Reality Magical Playbook:

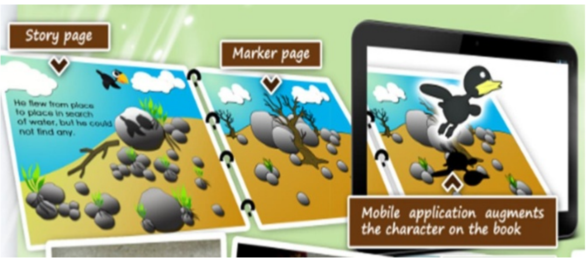

Story page
Marker page
He flew from place to place in search of water, but he could not find any.
Mobile application augments the character on the book

Fig. 2 General structure of the book [6]

The prototype concept of AR book presented in this paper is shown in Figure 2. The overall book design divided into story page and marker page. Story page covered the story line and illustration that illustrated the story within that current page. Marker page consists of marker that representing each 3D character with animation within that current page. This prototype is using handheld display (mobile device) for viewing the augmentation of the book because handheld display will help user to experience the AR concept while maintaining the context of reading normal book.

Natural interaction between user and the physical book should be included in AR book to maintain the context of normal user reading normal book. As AR book, it is aimed to enhance the traditional book, but not to replace the entire book [4]. Thus, the normal interaction with the book such as pointing will be presence in AR book. Tangible User Interface (TUI) in AR is the interaction that uses a physical environment to be tangible while interacting with the AR system [19]. In this prototype, finger can be used to interact with the AR book. Figure 3 shows the interaction with the AR book [6].

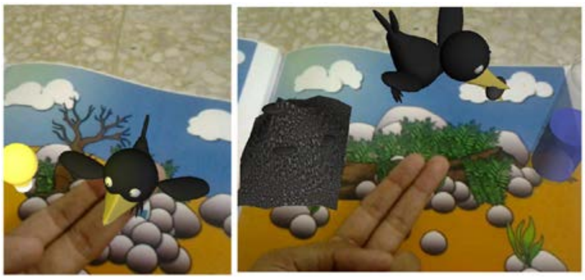


Fig 3 The interaction with the AR book using finger [6]

1.2.1 Mobile Application

Mobile AR application has gaining popularity nowadays due to mobile technology advancement. As mention early, mobile devices advanced in computing power and also in 3D graphic processing. This mobile application developed in Android platform. Figure 4 shows the mobile application installed in Android device.

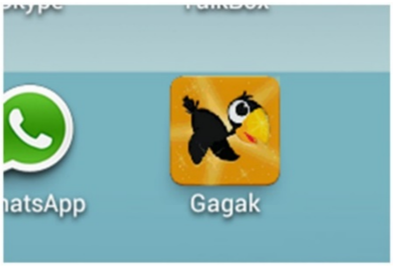


Fig. 4 Mobile application for AR book installed in Android device [6]

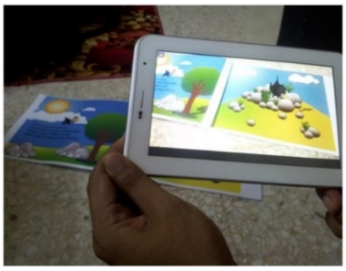


Fig. 5 The user uses the mobile application via mobile device to augment the character of the book[6]

The mobile application included with the ability to augment the 3D character and animation (the crow) with audio onto the marker page of the book. Figure 5 shows the user using the mobile application with the book.This application is not only augmenting the 3D character, animation and audio, but it also provides the users with narrator. The narrator is aimed to help children as guidance for them to read throughout the storyline. The learning number part is shows in figure 6.

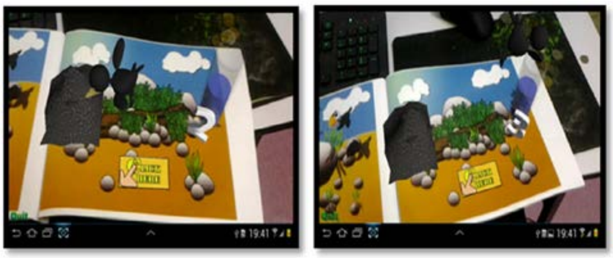


Fig 6 Learning number part[6]

Based on figure 6, learning number part is highly interactive designed for children. The user will interact with the book using their finger to count together with animation of the augmented 3D character on the book, leaving the natural means interaction between the user and the book. From here, the concept of engaging the student within learning environment in the learning process is fully applied[6].

1.3 Augmented Reality Children Storybook (ARCS)

The system will provide read aloud function so that children can listen and read along at the same time because much of the language children learn reflects the language and behaviour of the adult models they interact with and listen to [8]. Figure 7 shows the system architectural of the project [4]

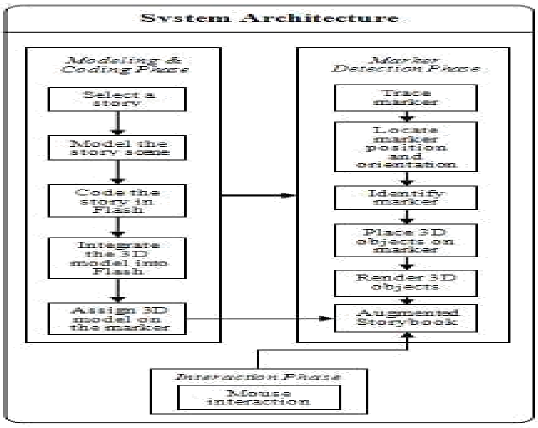


Fig. 7: System architecture graph for Augmented Reality Children Storybook (ARCS)[4]

As the targeting audience are the young learners who aged from 4 to 12 in Malaysia, the application have to be engaging and enjoyable so that it will motivate them. Hence, fun and attractive interface will help to capture the children's interest. However, every child has different learning skills and levels. Therefore, various stories will be included in the application for different categories of interest. The stories will include categories such as picture books, fiction, traditional literature which includes myths, legends and fairy tales and so on[4].

1.4 Children 's Interaction with Augmented Reality Storybooks

The story used for the prototype of the AR storybook was written in collaboration with two friends. The artefact in this study is a prototype of an AR storybook and the target group is eight-to ten-year-old Norwegian children. The AR book will be augmented using virtual 3D models, sound and interactive tasks. Sketching up a use case diagram, cf. Figure 8, where user activities as well as system responses are the main focus [9].

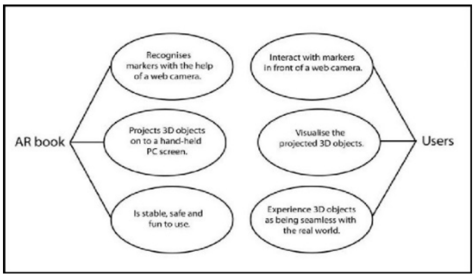

Recognises markers with the help of a web camera.
Interact with markers in front of a web camera.
AR book
Projects 3D objects on to a hand-held PC screen.
Visualise the projected 3D objects.
Users
Is stable, safe and fun to use.
Experience 3D objects as being seamless with the real world.

Fig. 8: Use case diagram of Children 's Interaction with Augmented Reality Storybooks [9].

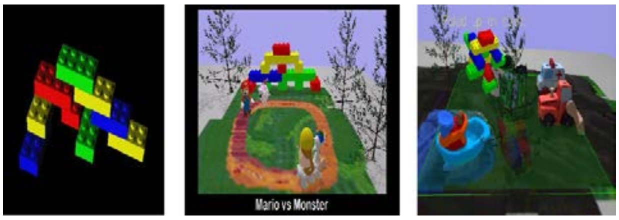
BuildAR Pro tracks and identifies markers in order to overlay the real world with virtual content. While writing the story, it had been decided that animals would be used for augmentation. Therefore, it was only natural that the pattern on the markers would also be animals, and markers were designed using more or less the same degree of detail [9].

1.1 Interactive Playspaces for Object Assembly and Digital Storytelling

A. Playspace

A playspace is an interactive system that aims to combine the advantages of doing a task physically and virtually. Figure 9 shows the novel active systems for virtual 3D content design applications – Block model assembly, Digital storytelling and 3D Scene design .In this setup, the user works on a planar work surface. The surface is divided into two parts – Play Area and Control Boxes. Any physical objects in the Play Area are tracked in real-time using the Kinect R color+depth camera.

The Play Area is exactly mapped to a part of the virtual world which is rendered on the display



Block Model Assembly

Digital Storytelling

3D Scene Design

Fig. 9: The novel interactive systems for virtual 3D content design

applications – Block model assembly, Digital storytelling and 3D Scene design [7].

Screen in front of the user. The tracked motion of physical objects is reflected in the virtual world in real-time. The Control Boxes can be used for gesture-based inputs [7]. Software framework of a playspace is shown in figure 10. The streams from the input modalities are given as input to the playspace algorithms – RGBD Processing module (for camera feed), Voice Recognition module (for microphone feed) and Event handlers for keyboard and mouse. The outputs of these algorithms are given as controls to the application running on the playspace. The application renders the virtual world and provide context-specific visual feedback on the display screen.

B. Assembly of Block Models.

Building block models with Lego R or Duplo R blocks is a popular hobby across adults and children. The block sets usually come with a set of instructions to put together a preconfigured model.

A system named DuploTrackis introduced, where a user works in a playspace with physical Duplo R blocks to build a pre-defined model. The system uses a novel 3D-tracking based guidance method to present instructions to the user. It also tracks the assembly process in real-time, points out any mistakes and helps correct them. The capability to track the assembly process also enables the system to learn how a new block model is assembled by a user. This learned representation can be used to share the model with other users via automatically generated representations like virtual 3D mesh models, static instructions, instruction videos or by boot-strapping it back into the system for guiding a new user [7].

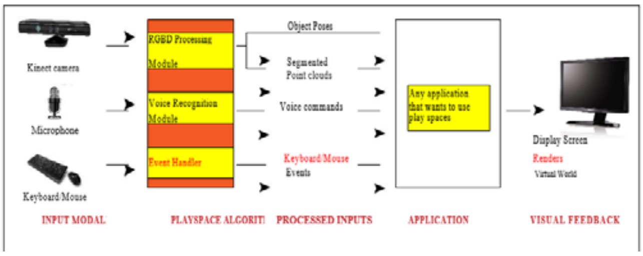


Fig. 10: Software framework of a playspace

C. Digital storytelling:

The system allows a user to act out a story using rigid puppets and automatically converts that into an animation. Further, it also allows the user to record multiple takes for the same story and merge them automatically after the user has roughly annotated them based on his liking. This is helpful when the user wants to try out different styles and later merge them [7].

Fig.11(a) Toys for storytelling. for story telling

Fig.11(b) 3D-Puppetry system

Figure 11: Natural and intuitive interfaces for storytelling. (a) Toys and puppets are the traditional ways of natural story telling. (b) The 3D-Puppetry system tracks the moving physical objects using a Kinect R camera and renders their tracked virtual replicas on the

An intuitive interface to tell a visual story for novice users is through physical puppets and toys (Figure 11.a). Hence we can develop systems which automatically track and transfer the acted out motion to virtual characters and hence record an animation.

The 3D-Puppetry system uses the framework of a playspace. Figure 11.b shows a user using the system. As is the case with playspace framework, the user first scans in the physical objects that he intends to use in the story. Then he acts out the story using objects in the Play Area which the system tracks in real time and renders replicas in a pre-selected virtual environments on the display screen in front of the user. This rendering is also recorded as a video which is the resulting animation. This system allows user to use some keyboard and mouse-based controls to edit the animation later by changing light positions, camera viewpoint etc [7].

II. Multimodal, Multi-User and Adaptive Interaction for Interactive Storytelling

The design of multimodal, multi-user, and adaptive interaction model follows some requisites for the design of multimodal interfaces described by Reeves et al. [2004]. Adapting some of these concepts to the interactive storytelling domain, defines the following requisites to the interaction system [8]:

Natural Interaction: The multimodal interaction must be natural. The viewers must feel comfortable interacting with the system;

Adaptable Interaction: The multimodal interface must adapt itself to the needs and abilities of different viewers;

Consistent Interaction: The result of an input shared by different interaction modalities must be the same;

Error Handling: The system must prevent and handle possible mistakes in the interaction, as well allowing the viewers to easily undo their actions;

Feedback: The system always must give a feedback to the viewer's when some action resultant from a multimodal interaction be executed;

Equal Interaction: In a multi-user scenario,the interaction system must offer equal possibilities of interaction to all viewers [8].

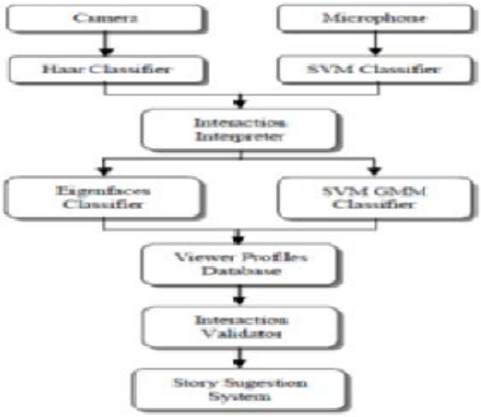


Fig. 12 Multimodal, Multi-User and Adaptive Interaction architecture [8]

The multimodal interaction interface is based on gestures and speech. The choice of these interaction modalities was made due to the need of natural interaction modalities in a multi-user setting. Gestures and speech provide a natural interaction interface and allow the interaction of several users by using computer vision and speech recoginition technique. The viewers are free to use both interaction modalities.

The architecture of the interaction system presented in this paper as show in figure 12. The system uses a conventional camera and a microphone to capture the input of the system. The viewers are located on the video input by the HaarClasifieralogrithm, and the viewer's speech is recogized by the SVM Classifier based on the audio input. The interaction Interpreter module analyses and interprets and the viewer's gestures and speech commands. Next, the EigenfaceClassifer and the SVM GMM Classifier identify the viewer based on the profile of the viewers (which is stored in the Viewers Profiles Database). Each interaction is the recorded in the appropriated viewer's profile. The profile management updates the viewer's profile based on the viewer's interactions and the atmosphere traits associated to the events as modeled in the Atmosphere Database. Before the viewer's interaction affects the system, the Interaction Validator module checks if the viewer is not interacting for the second time in the same option (for example to avoid a viewer voting more than one time in the same option). Finally, the user interaction is sent to the Story Suggestion System [8].

III. *Location-based Storytelling in the Urban Environment*

The user experience and user interaction with the system was designed using sketching, mockups, and paper prototypes in parallel with the story writing activity. We wanted to combine the digital elements of the story with tangible interactions with real-world locations and objects such as inscriptions and symbols on buildings, paper maps, and physical props similar to what might be found in a theatre production of the story. Other props explored included marked envelopeswith physical evidence or clues to be opened at particular points of the story, and to be passed between the users. Another approach explored to tie the digital experience closely to the physical surroundings was the use of printed photographs of actual locations in the user's current surroundings overlaid with fictional characters and objects from the storyline [10]. Technical set-up of the prototype is shown in figure 13.

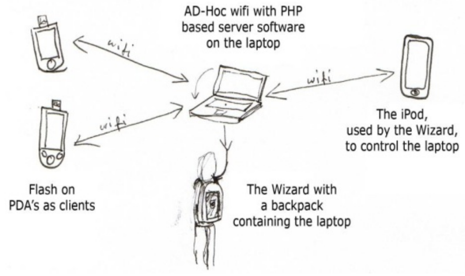


Fig. 13 Technical set-up of the prototype.[10]

Proceeding from one scene to the next

Once the two detectives have gathered enough pieces of information at a particular location they are prompted to move on to the next scene at a different place. Rather than providing way-finding information on the PDAs for this purpose, the users are provided with a physical map of the city with key locations of the story highlighted. However, in order to keep the path through the city flexible and secret, each location is annotated with a unique symbol rather than numbers or letters, making it impossible for the detectives to know where to go next. Increasing the challenge, the correct symbol can only be obtained through collaboration. Based on the information gathered from individual interrogations, each detective will at some point in time be provided with half the symbol needed to move on. Only when both halves have been obtained is it possible for the two participants to work outgo to next by locating the corresponding composite symbol on the physical map (figure 14) [10].

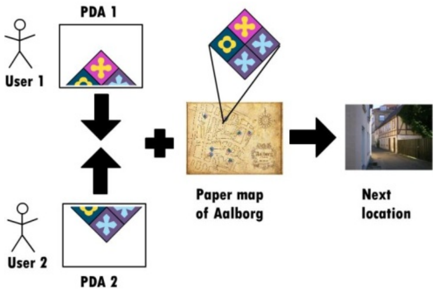


Fig. 14. Finding the next location of the story from a composite symbol and a physical map [10].

IV. *Storytelling in Collaborative Augmented Reality Environments*

Figure 15 shows the architecture of the system in regard to the several modules and layers used. Every layer is parted in several AI sub-modules to improve the possible evolution of the systems abilities, as we wanted to design a reusable software system with the possibility to replacemodules in regard to project specifications and scientific research ideas [11].

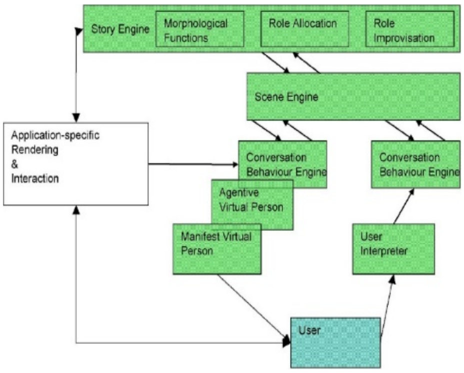


Fig. 15: Architecture of CSCIS system[11]

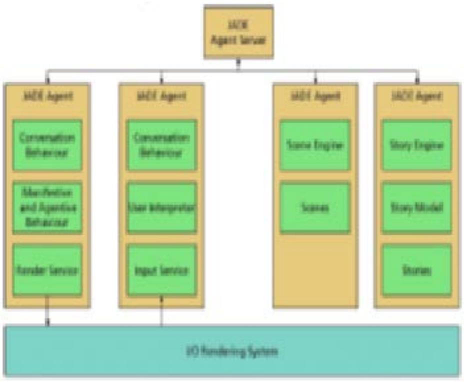


Fig. 16: Development of the CSCIS System[11]

Figure 16 provides a sketch of the implementation of the system. We used several AI- related software packages to develop the story engine (done with Prolog), as well as the conversational behavior, agentive and manifest modules of the several agents, with this agents playing roles (virtual characters) in AR story environment (done with Jess, the

Java Expert System shell [Fri01]. Communication between the several modules is done using the JADE Agent Platform [Bel01].

The authoring process of Interactive Stories is supported in regard to the definition of scenes for the AR environment and the relation of scenes to the morphological story model (functions and roles), as well as to improvisational features: see Schnieder [Sch 02] [11].

V. Digilog book for temple bell tolling experience based on interactive augmented reality

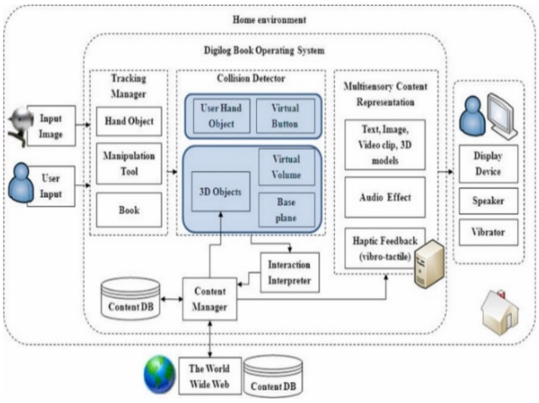


Fig. 17 Dialog Book system architecture[12]

Fig. 17 shows the Dialog Book system architecture based on the proposed interactive AR system. Based on input images from a camera, a computer vision based tracking manager recognizes and tracks a paper book, a manipulation tool, and a hand object. The collision detector then inspects penetrations between a virtual line created by the manipulation tool and a bounding volume of the augmented 3D objects that are based on the book. The detector also checks an occulded area between the virtual buttons and the user's hand objects. At this point, the interaction interpreter conducts examinations like 3D object pointing and hitting, movement interactions or hand interactions for pushing virtual buttons[12].

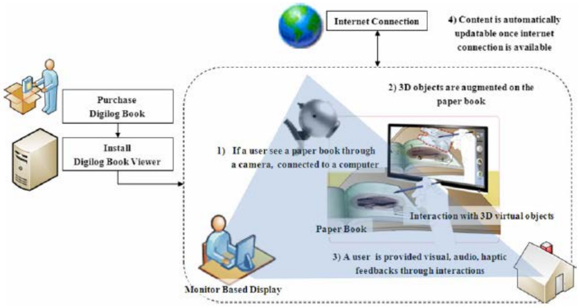

Internet Connection
4) Content is automatically updatable once internet connection is available
Purchase Digilog Book
Install Digilog Book Viewer
2) 3D objects are augmented on the paper book
1) If a user see a paper book through a camera, connected to a computer
Interaction with 3D virtual objects
Paper Book
3) A user is provided visual, audio, haptic feedbacks through interactions
Monitor Based Display

Fig. 18 Conceptual figure of Dialog book usage [12]

Next, the content manager composes proper multisensory content in order to react to the user input. The multisensory content consists of visual feedback (text, images, background sounds) and haptic feedback (vibro-tactile interactions). Finally, a display device, a speaker, and a vibrator of the manipulation tool represents the multisensory content. Additionally, if AR content in a remote database is updated, then the AR content in a remote database is updated, then the AR content of a Dialog Book will be updatable through an Internet connection [12]. Fig 18 shows conceptual figure of Dialog book usage.

VI. User Interaction in Mixed Reality Interactive Storytelling

The approach used in [13] is character-based, which means that the narrative is driven by the individual roles of each of the virtual actors, rather than by an explicit plot representation. The actors' roles are formalised as plans, which are executed in real time using a modified Hierarchical-Task Network planning algorithm. During execution, the planner selects the next action for a character, this action being played in the virtual environment, which is also updated to take into account its consequences. When an action fails (i.e. its intended outcome is not achieved), another course of action is generating through re-planning. The real-time selection of action supports interactivity, as the user can interfere with the environment, changing the executability conditions of potential actions[13]. The system overview of user interaction is shown in figure 19.

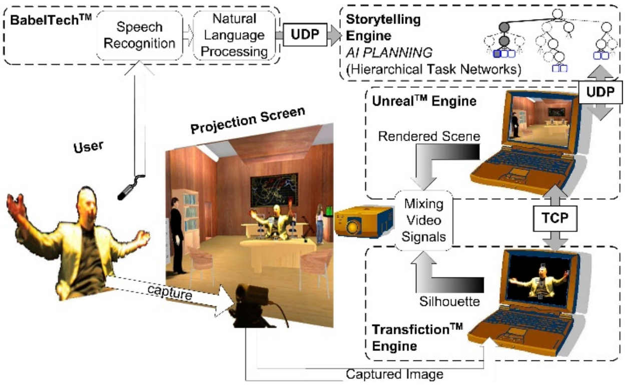


Fig. 19 System overview of User Interaction in Mixed Reality Interactive Storytelling [13].

VII. REVIEW AND CONCLUSIONS

The combination of mobile augmented reality and interactive storytelling can be used in various applications. The Interactive augmented reality storybook surveyed in this paper are available both as physical book and AR book [5][6]. It can also be used in Mobile phones (PDA and smart phones)[4][6] and [5] can be used in desktop and mobile phones. With reference to markers, [4] author as used visible, black bordered markers whereas, work [5][6] uses invisible markers which is one of the added advantage to the application. The basic difference between all the surveyed paper lies in the different user interaction. [4] Uses mouse interaction, [6] uses tangible user interface (finger), [8] uses multimodal interactions (gesture and speech), [12] uses hitting, movement interaction or hand interaction. The real-time selection of action supports interactivity, as the user can interfere with the environment changing the executable condition of potential actions [13]. Considering the story engine, several AI related software package are used to develop the story engine in [11]. Some of the other features are multi modal (gesture and speech), multi-user, Adaptive Interaction (based on viewer's option), location based, Collaborative and Interactive options.

Education can be much more interesting and interactive by applying computer technologies such as multimedia into in. In order to promote the reading habit to the children nowadays, the Interactive AR Storybooks not only provides knowledge but entertainment at the same time. With so many successful examples of how computer technologies were applied in education, this Interactive Storybooks based on Mobile augmented reality will be one of it as well. The

Augmented reality storybooks are used to learn either English, learning numbers using an old folklore literature or any other subject. Since the children prefer audio and graphics, the Augmented Reality Storybook will provide not only these but allow interaction so that children can learn andplay a role in the story at the same time.

REFERENCE

[1] Zhanpenghuang, pan hui, christophpeylo, dimitrischatzopoulos “Mobile augmented reality survey: a bottom-up approach”, arxiv:1309.4413v2 [cs.gr] 18 sep 2013

[2] Veronica teichrieb, joaopaulosilva do monte lima, eduardolourenc¸oapolinario, thiagosoutomaiorcordeiro de fariasmarcioaugustosilvabueno, judithkelner, and ismael h. F. Santos, “A survey of online monocular markerless augmented reality”, International journal of modeling and simulation for the petroleum industry, vol. 1, no. 1, august 2007

[3] Edirleisoares de lima, brunofeijó, simonebarbos , fabioguilherme da silva, antonio l. Furtado, cesar t. Pozzer , angelo e. M. Ciarlini , “multimodal, multi-user and adaptive interaction for interactive storytelling applications” , sbc - proceedings of sbgames 2011”Behrangparhizkar, tan yi shin, arashhabibilashkari, yap sing nian, “Augmented reality children storybook (ARCS) “, 2011 international conference on future information technology ipcsit vol.13 (2011) ©
(2011) iacsit press, singapore

[4] Albertinadias, “Technology enhanced learning and augmented reality:

[5] An application on multimedia interactive books”,International business & economics review, vol.1, n.1 issn 1647-1989

[6] Azfar bin tomi, dayangrohayaawangrambli,”An interactive mobile augmented reality magical playbook: learning number with the thirsty crow“, 2013 international conference on virtual and augmented reality in education doi:10.1016/j.procs.2013.11.015

[7] Ankitgupta , “Interactive playspaces for object assembly and digital storytelling”, university of washington 2013

[8] Edirleisoares de lima, brunofeijósimonebarbosa ,fabioguilherme da silva, antonio l. Furtado, cesar t. Pozzer,angelo e. M. Ciarlini, “multimodal, multi-user and adaptive interaction for interactive storytelling applications”, sbc - proceedings of sbgames 2011

[9] Olaugeiksund, “children’s interaction with augmented reality storybooks -a human-computer interaction study”, spring 2012

[10] Jenipaay and jesperkjeldskov , anderschristensen, andreasibsen, dan jensen, glen nielsen and renévutborg, “ location-based storytelling in theurban environment”, OZCHI 2008 Proceedings ISBN: 0-9803063-4-5

[11] Norbert braun, “storytelling in collaborative augmented reality environments” , WSCG'2003,Feburary 3-7 , 2003, Plzen Czeh Republic

[12] Taejin Ha, Youngho Lee, Woontack Woo,"Digilog book for temple belltolling experience based on interactive augmented reality", Virtual Reality (2011) 15:295–309 DOI 10.1007/s10055-010-0164-8

[13] Cavazza, M. O. et. al. (2003) 'User interaction in mixed reality interactive storytelling', 2nd IEEE/ACM international symposium onmixed and augmented reality in Proceedings of the 2nd IEEE/ACM international symposium on mixed and augmented reality.Washington: IEEE, p.304.

[14] Isbell, R.,” Telling and retelling stories – learning language andliteracy”, Young Children, 2002, 57(2), 26–30.

Er. Mrs. Sagaya Aurelia(November 9,1978) par-time research scholar in Bharathidasan university . Now she is with department of Computer Science, Faculty of Education, Azzaytuna University, Bani-walid, Libya. She received her Diploma in Electronics and Communication (1997),B.E (Bachelor of Engineering specialized in Electronics and Communication Engineering(2000) and M.Tech in Information Technology(2004),she has alsodoneherPostgraduation diplomas inBusiness Administration(PGDBA) and Journalism and Mass Communication(PGDJMC). She has receivedBrainbench certification in HTML. Her current research interest includes Virtual reality, augmented reality and Human Computer Interaction and User interface Design. She has authored14 papers and attendance several national and international level workshops and conferences.

Dr. Durai Raj is currently working as Assistant Professor, School of Computer Science, Engineering and Applications, Bharathidasan University, Tiruchirappalli, Tamilnadu, India. He completed his Ph.D. in Computer Science as a full time research scholar at Bharathidasan University on April, 2011. He received master degree (M.C.A.) in 1997 and bachelor degree (B.Sc. in Computer Science) in 1993 from Bharathidasan University. Prior to this assignment of Assistant Professor in Computer Science at Bharathidasan University, he was working as a Research Associate at National Research Centre on Rapeseed-Mustard (Indian Council of Agricultural Research), Rajasthan, and as a Technical Officer (Computer Science) at National Institute ofAnimal Nutrition and Physiology (ICAR), Bangalore for 12 years. He has published 26 research papers in both national and international journals. His areas of interest include Artificial Neural Network, Soft Computing, Rough Set Theory and Data Mining.

Dr. Omer Saleh MahmodJamah (January 25,1973) is now the Director of Post graduate cum Research and Development and Head of the department of Computer science, Faculty of education, Azzaytuna university, Baniwalid, Libya. He received his B.Sc. in Control System and Measurement (1995), M.Sc. in Electrical and Computer Measurement (2004), and Ph.D. in Electrical engineering, Automatics computer science and electronics from AGH University of technology, Krakow, Poland. He has done his Diploma in Planning and time management from Canada Global Centre, Canada. Now he is heading Computer Science department, Faculty of Education, Azzaytuna University, Baniwalid, Libya. His research interest includes multicriteria optimization for solving optimal control problems and Fuzzy logic. He has published 12 papers and attended various national and international Level conferences and workshops.

Architecture of a Multi Agent Intelligent Decision Support System for Intensive Care

Pedro Gago, Manuel Filipe Santos

***Abstract*—**Intensive Care Units are fertile ground for Decision Support Systems (DSS). Not only there is a perceived need for better tools but also the number of devices present greatly reduces the need for manual data entry and thus potentially makes the system much less disruptive of the physicians' routines. Moreover, designing DSSs as Multi Agent Systems (MAS) allows for "nicer" architectures and for conceptual gains as each functionality can be abstracted and each component of the whole system can be tackled individually. In this paper we present the architecture of the INTCare System, a Multi Agent Intelligent DSS designed to work in an ICU setting. INTCare architecture allows it to manage several prediction models for the same prediction objective deciding in real time which to use when a prediction is required. Moreover, included in the system are rules that allow it to monitor its predictive performance and to adjust its ensemble based prediction models as to reflect that performance.

***Keywords*—**decision support, intensive care, multi agent systems.

I. Introduction

RECOMENDATIONS on health related Decision Support Systems (DSS) development stress the importance of making such systems part of the intended users existing routine. There are indications that user resistance is very high whenever extensive and time consuming data entry is required in order to me able to get an answer from the system. Successful DSS try to keep manual data entry to a minimum by trying to connect to available data sources and reduce manual data entry [1], [2].

In an ICU with an Electronic Health Record system in place data for decision support is already in digital format and for seamless integration in the existing workflow a DSS must only collect and use it. Furthermore, real time data is collected from each patient by means of the bedside monitors that are ever present in such places creating new opportunities for real time data analysis that can potentially lead to new forms of decision support in the ICU.

Making use of all that data makes it possible to use supervised learning algorithms in order to build prediction models for the patient's final outcome and for a number of other intermediate outcomes of interest (e.g. predicting organ failure or length of stay in the ICU). Real time data from bedside sensors is likely to allow the design and creation of new tools that may help prevent degradation of the patient's conditions (e.g. alert for possible organ failure). [3]

A Knowledge Discovery from Databases (KDD) approach may be used to provide a guide from data collection all the way through model evaluation and by using it we were able to get some initial results. Keeping in line with the low intrusiveness objective, we designed a system that we called INTCare that aims at being autonomous in that it operates and keeps itself up to date with little or no human intervention. Subsequent iterations of through the KDD cycle must require no human intervention and the system is designed in such a way as to automate the process. Moreover, as the stay in ICU progresses, more and more data becomes available for each patient. In fact, not only analysis and other exams are performed but also bedside monitor are continuously supplying new information about the patient. Despite the fact that good predictions for final outcome can be made with data from the first 24 hours in the ICU, it is to be expected that better results be achieved as more and more information is considered. An autonomous DSS should use whatever prediction model gives the best results, automatically using the best model given the available data.

The Multi Agent System (MAS) architecture for such a system emerges naturally as the KDD process is divided in sub steps and each of them has clear entry and exit products. INTCare has an agent for data preprocessing, another for choosing the most adequate model for each situation, another for performance evaluation, and so on. Even if each agent has limited capabilities and responsibilities, together the agents contribute to the system's success.

In the next section we present case for DSS in the ICU and in section III we present MAS. Next, in section four the INTCare System is presented. Finally, in section five, we discuss the relevance of the system.

II. Decision Support for Intensive Care

Decision Support Systems are especially adequate for those situations where the wealth of data to be considered exceeds human capacity and/or where there are no absolute theoretically sound answers to the problems to be addressed. Multi objective problems are particularly difficult to humans

This work has been supported by FCT – Fundação para a Ciência e Tecnologia in the scope of the project: PEst-OE/EEI/UI0319/2014.

Pedro Gago is with School of Technology and Management of the Polytechnic Institute of Leiria, Leiria, Portugal (phone: 351 244820300 email: pedro.gago@ipleiria.pt).

Filipe Portela, José Machado, António Abelha and Manuel Filipe Santos are with the Algoritmi Research Centre, University of Minho, Guimarães, Portugal (email: cfp@dsi.uminho.pt, jmac@di.uminho.pt, abelha@di.uminho.pt, mfs@dsi.uminho.pt).

due to the number of variables to be considered. Intensive Care Units are particularly well suited for the implementation of DSS as they present those characteristics of high data volume and dimensionality and multi objective. In fact, even in an ICU, or particularly in an ICU tradeoffs are constantly being made when choosing treatment strategies. Factors considered include cost, risk, impact on other health aspects, and so on.

A. Past experiments

In past work we performed some experiments in an ICU setting, where we now have automated the data acquisition process and created automatic detection methods for potentially relevant clinical situations (out of range events for important monitored variables) [6],[7]. Presently, relevant data is available in digital format including admission data, exam results, bedside monitor variables, medication and procedures performed.

Using that data, prediction models were built and encouraging results were achieved. However, Decision Support for ICU calls for seamless integration with the existing workflow and for a constant use of the most up to date information that demands. With that goal in mind we successfully investigated the possibility of using ensembles of prediction models in order to improve our results. Furthermore, a mechanism for automatic adjustments to the ensembles was designed and tested with encouraging results [8].

B. KDD in decision support

The interest in Knowledge Discovery from Databases (KDD) and Data Mining (DM) arose due to the rapid emergence of electronic data management methods. In 1997, the Gartner Group suggested that DM would be one of the top five key technologies that would have a major impact in the industry within the next years. In effect, these techniques are now widespread several examples can be found in areas so diverse as marketing, banking, manufacturing and production, brokerage and securities trading or health care [9]-[11].

Within the Medicine arena, huge databases, with large, complex and multi-source information (e.g. text, images or numerical data), are commonplace. However, human experts are limited and may overlook important details. Furthermore, the classical data analysis (e.g. logistic regression) breaks down when such vast amounts of data are present. Hence, an alternative is to use automated discovery tools to analyse the raw data and extract high level information for the decision-maker [12].

Some of the most used Machine Learning algorithms are Decision Trees [13], [14] and Artificial Neural Networks [15], [16]. Some work has been developed on the use of ensembles for supervised learning, where a set of classification/regression models are combined in some way to produce an answer [17].

III. MULTI AGENT SYSTEMS

The term agent is a metaphor allowing various definitions, interpretations and taxonomies. Actually, no one of them is universally accepted, despite this some positions are considered referential [18]-[20]. One of the most comprehensive definitions of agent was proffered by Jennings et al., and is based on a strong and weak view [18]. Applications of agents are widespread and can be found in travel planning [21],[22] and even as components of DMSSs [23].

In the context of this work, the AIMA definition prosecuted by [24] was adopted, stating that an agent is an entity capable of perceiving the environment and actuating on that environment. From a software engineering point of view, an agent is an abstraction that allows the construction of more complex systems designated by Agent-Based Systems, Agencies or Multi Agent Systems.

The requisites inherent to a situated and active system, in addition to the previous experience accumulated in the conception of agencies such as the DICE system [25], and the AIDA agency [26], dictated the adoption of the agent technology in the development of INTCare.

Next, after a formal definition of the MAS that composes INTCare, the various types of agents integrating the system are presented and discussed.

IV. INTCARE SYSTEM

Autonomy is an important feature for a DSS, even more so for a DSS in a clinical setting. Such independence from human experts makes it possible for the system to work for a long time and to blend into the background, increasing the likelihood of adoption by the physicians. Moreover, given the wealth of data continuously being produced in an ICU, the system may be built in a way that profits from it. It is likely that predictions will be asked in different times of the patient's stay in the ICU. In fact, a prediction may be requested after the first 24 hours in the ICU and another after a week of stay. It is to be expected that a prediction made with the data of the first few days in ICU will be more accurate than another that uses data from the first 24 hours of stay. In order to account for the fact that data is continuously being gathered for each patient, INTCare may use several prediction models for each required outcome. When a prediction is needed, the system uses the most adequate prediction model given the available data. Additionally, given the possibility of collecting outcome data, INTCare is capable of verifying its predictions accuracy. In an effort to achieve better results it may resort to ensembles of prediction models and use a dynamic weighted majority voting strategy that has shown to perform well in past experiments with medical data. [27].

A. Formal definition

Conceptually, the INTCare system can be viewed as set of four subsystems: Data Entry, Knowledge Management, Inference and Interface. In more formal terms, the INTCare system is defined as a tuple $\Xi \equiv \langle C_{INTCare}, \Delta_{INTCare}, a_{init}, a_{pp}, a_{inf}, a_{ct}, a_{pf}, a_{pred}, a_{ens} \rangle$, where:

$C_{INTCare}$ is the context and corresponds to a logical theory, represented as a triple $\langle Lg, Ax, \Delta \rangle$, where *Lg* stands for an extension to the language of programming logic, *Ax* is a set of axioms over *Lg*, and Δ is a set of inference rules;

$\Delta_{INTCare}$ is the set of bridge rules defining the interaction among the systems' components (the agents);

$A_{init}, \ldots, a_{ens}$ are the system's agents.

This formalism corresponds to a logical framework, suitable to specify agent-oriented systems based on the notion of context logic, and some properties of object-oriented design such abstraction, encapsulation, modularity and hierarchy (Santos 1999). In this work, the agents are represented as logical theories with a specific context (different agents may involve different contexts). Several agents (i.e. contexts) can be put together and be able to reason about the behaviour of the entire system as a (heterogeneous) logical theory. A set of special rules called *bridge-rules* is applied to provide the interface among agents and systems of agents. These rules describe the agents' reactions to events occurring in their environment. The agents include a set of event types, and a set of time points. Next, the overall system is described making use of this formalism, explaining it in some detail.

B. *Initialization Agent*

The first run requires human intervention. At that time, prediction objectives are set and input data is defined. Moreover, the whole process is fine-tuned and the first prediction models are created. The Initialization Agent provides an interface for the human expert to specify those parameters that will be needed for the automatic operation of the system. It is through this agent that input data for each prediction model is defined and that model ordering is established indicating the model to be used when data collected is sufficient for more than one. In addition, the systems allows for the use of *ensembles* of prediction models and that must also be specified.

This agent is characterized by the events described in Table i:

Table i - Events for the Initialization Agent

Event	Description
set_preferences	indicates relative preference for models regarding the same prediction objective. Populates a modelList ordered from best to worst model. When a prediction is required, the system uses the best model that has enough available data.
set_type	defines the type of model (single model or ensemble of models)
define_inputs	defines the inputs required for the prediction model
define_range	defines the admissible range for each attribute

C. *PreProcessing Agent*

This agent's responsibility is to prepare data to be used both when requesting a prediction and as training data for the creation of new prediction models. It connects to the different data sources and populates the adequate data tables with patient data, as it becomes available. Moreover, this agent also strives to guarantee data quality by filtering values that are obviously incorrect (as defined using the Initialization Agent).

In Table ii we present the events for this agent:

Table ii - Events for the PreProcessing Agent

Event	Description
get_data	retrieves the relevant new data from the specified data source
evaluate	checks datafor out of range values.
update	updates the tables where the data must be stored

D. *Inference Agent*

As the patient's stay in the ICU progresses more and more data is collected. If in the first day of stay predictions must be made based on relatively scarce data after some days in the ICU, there possibly is enough data for more accurate predictions, using different prediction models.

Whenever a prediction is requested this agent confirms the data available for the current patient and, using the list populated by the initialization model, selects the most adequate prediction model in order to get a prediction.

This agent is characterized by the events presented in Table iii:

Table iii - Events for the Prediction Agent

Event	Description
get_model	retrieves the most adequate prediction model given the available data and the prediction goal.
apply_model	uses the available data and the selected model to get a prediction
send_result	sends the prediction result to the requester agent
update_dw	saves the value predicted by each model (that will be used by the Performance Agent when the real value becomes available)

E. *Control Agent*

Whenever new data is made available to the system it is necessary to decide if it is relevant or if it should be discarded. Relevant data corresponds to data defined using the Initialization Agent to be used as inputs for the prediction models or data that corresponds to the prediction goals and thus can be used to verify the predictions made (eg. Discharge data that is used to confirm outcome prediction results.). This agent then routes the data to the appropriate agent.

This agent is characterized by the operations presented in Table iv:

Table iv - Events for the Control Agent

Event	Description
scan	scans all available data sources for new data
evaluate_input	checks if the data is relevant for a prediction model as input data
evaluate_output	checks if the data is relevant for a prediction model as output data
message_pp	sends a message to the PreProcessing Agent with the relevant input data
message_per	sends a message to the Performance Agent with the relevant output data

F. Performance Agent

INTCare's autonomy relies on its ability to update the prediction models used. In fact, performance data is used not only to alter the relative weights inside model ensembles but also serves as basis for model deletion or new model creation. For a number of prediction objectives the normal sequence of events in an ICU will, sooner or later, make available data that will confirm or disprove the predictions made. When that happens, INTCare will update the weights of the models in the *ensembles* punishing those with missed predictions and rewarding the most accurate ones. Moreover, in order to maintain a good performance, low performing prediction models are removed from the ensemble eventually being substituted by new ones, created using the most recent data as training data.

Table v contains the events for this agent:

Table v - Events for the Performance Agent

Event	Description
update_weights	updates the weights of the models in the ensemble, given their predictions and the new data that confirms or counters those predictions
remove_model	removes from the ensemble those models with weights bellow a predefined threshold
train_model	trains a new prediction model for the present goal and using the most recent data.
add_model	adds a new model to the ensemble

G. Interface Agent

Physicians use the system through the Interface Agent that gathers their requests, routes them to the appropriate agents and finally presents the outcome back to the requesters. User interface must friendly and the overall operation of the system should be integrated in usual medical practice in a seamless manner.

Table vi - Events for the Interface Agent

Event	Description
pred_request	Requests a prediction from the *Inference Agent* for the specified outcome and patient
pred_present	Presents the predicted result to the physician

H. Bridge Rules

The architecture is completed by means of the bridge rules in Table vii. This formalism provides us with a clear way to describe the interaction between the agents.

Table vii - INTCare's bridge rules

	Bridge-Rule	**Description**
1	$\frac{a_{int}: occurs(re\_prediction, t)}{a_{pd}:[occurs(get\_models, t) \wedge occurs(predict, t) \wedge occurs(send\_data, t) \wedge occurs(update\_dw, t)]}$	When the *Interface* agent requests a forecast, the *Prediction* agent gets the adequate model from the Knowledge Base applies it and finally sends the results back to the *Interface* agent and updates the Data Warehouse with the prediction results (to be used later by the Performance Agent).
2	$\frac{a_{ct}: occurs(message\_pp, t)}{a_{pp}:[occurs(get\_data, t) \wedge occurs(evaluate, t) \wedge occurs(update, t)]}$	When the Control Agent messages the Pre Processing Agent indicating that new input data is available the Pre Processing Agent gets the data, checks it for out of range values and stores it in the adequate tables for later use.
3	$\frac{a_{ct}: occurs(message\_per, t)}{a_{pp}:[occurs(get\_models, t) \wedge occurs(update\_weights, t)]}$	When the Control Agent messages the Performance Agent indicating that new output data is available the Performance Agent: gets the models that were used to predict that variable and updates the weights of the models in those ensembles.

V.DISCUTION AND CONCLUSIONS

Agents provide a convenient abstraction to deal with complex problems. INTCare is a MAS that aims at being autonomous and capable of adjusting to its environment either by choosing the most adequate prediction models depending on the available data or by using a weight adjustment strategy for changing the models weights inside the ensembles in order to maintain a good prediction performance over time.

VI. ACKNOWLEDGMENTS

This work has been supported by FCT – Fundação para a Ciência e Tecnologia in the scope of the project: Pest-OE/EEI/UI0319/2014.

The authors would like to thank FCT (Foundation of Science and Technology, Portugal) for the financial support through the contract PTDC/EEI-SII/1302/2012 (INTCare II).

REFERENCES

[1] E.S. Berner and T. J. La Lande, “Overview of Clinical Decision Support Systems”, in E. S. Berner (ed.), “Clinical Decision Support Systems – Theory and Practice”, 2nd Edition, Springer, pp 3-22, 2007.

[2] R. Vahidov, G. Kersten, “Decision station: situating decision support systems”. Decision Support Systems, vol. 38, no. 2, pp. 283-303, 2004.

[3] F. Portela, M. F. Santos, J. Machado, A. Abelha, A. Silva, F. Rua. “Adoption of Pervasive Intelligent Information Systems in Intensive Medicine”. Procedia Technology. Volume 9 - HCIST, pp. 1022-1032. Elsevier, 2013.

[4] F. Portela, P. Gago, M. F. Santos, J. Machado, A. Abelha, A. Silva, F. Rua, C. Quintas and F. Pinto, “Intelligent and Real Time Data Acquisition and Evaluation to Determine Critical Events in Intensive Medicine”. Procedia Technology. Volume 5 - HCIST, pp. 716-724. Elsevier, 2012.

[5] F. Portela, A. Cabral, A. Abelha, M. Salazar, C. Quintas, J. Machado, J. Neves, M. F. Santos, “Knowledge Acquisition Process for Intelligent Decision Support in Critical Health Care”. IGI-Global book on Information Systems and Technologies for Enhancing Health and Social Care. pp. 55-68. ISBN13: 9781466636675. IGI Global, 2013.

[6] P. Gago, M. F. Santos , “Towards an Intelligent Decision Support System for Intensive Care Units”. Presented at the Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications, 18th European Conference on Artificial Intelligence, Patras, Greece, 2008..

[7] E. Turban, J. Aronson, T. Liang, “Decision Support Systems and Intelligent Systems”, Prentice Hall, New Jersey, 2004.

[8] M. F. Santos and C. Azevedo, “Data Mining – Descoberta de Conhecimento em Bases de Dados”, FCA Editora, Lisbon (in Portuguese), 2005.

[9] S-h. Liao, “Knowledge management technologies and applications-literature review from 1995 to 2002”. Expert Systems with Applications, vol. 25, no. 2, pp. 155-164, 2003.

[10] D. Hand, H. Mannila, P. Smyth, “Principles of Data Mining”. MIT Press, Cambridge, MA, 2001.

[11] J. Quinlan, “Induction of Decision Trees”. Machine Learning, vol. 1, no.1, pp. 81-106, 1986.

[12] N. Lavrac, E. Keravnou, B. Zupan, “Intelligent Data Analysis in Medicine and Pharmacology”. Kluwer, Boston, 1997.

[13] S. Haykin, “Neural Networks and Learning Machines” (3rd Edition). Prentice-Hall, New Jersey, 2008.

[14] R. Dybowski, “Neural Computation in Medicine: Perspectives and Prospects”. In: Malmgreen et al. (eds) Proceedings of the ANNIMAB-1 Conference (Artificial Neural Networks in Medicine and Biology). Springer, Berlin Heidelberg New York, pp 26-36, 2000.

[15] T. Dietterich, “Ensemble methods in machine learning”. In: Kittler and Roli (eds), Multiple Classifier Systems, LNCS 1857, Springer, pp 1-15, 2001.

[16] R. Jennings, J. Wolldridge, “Agent Technology Foundations, Applications and Markets”, Springer, Berlin Heidelberg New York, 1998.

[17] G. Weiss, “Multiagent Systems - A Modern Approach to Distributed Artificial Intelligence”, MIT Press, Cambridge MA, 1999.

[18] J. Ferber, “Multi-Agent Systems – An Introduction to Distributed Artificial Intelligence”, Addison-Wesley, 1999.

[19] H. S. Yim, H. J. Ahn, J. W. Kim, S.J. Park, “Agent-based adaptive travel planning system in peak seasons”. Expert Systems with Applications, vol. 27, no. 2, pp. 211-222, 2004.

[20] A. M. Garcia-Serrano, P. Martinez, J. Z. Hernandez, “Using AI techniques to support advanced interaction capabilities in a virtual assistant for e-commerce”. Expert Systems with Applications, vol. 26, no. 3, pp. 413-426, 2004.

[21] T. J. Hess, L.P. Rees, T. R. Rakes, “Using autonomous software agents to create next generation of decision support systems”. Decision Sciences, vol. 31, no. 1, pp 1-31, 2000.

[22] S. Russel and P. Norvig, “Artificial Intelligence – A Modern Approach” 3rd Edition. Prentice Hall, New Jersey, 2010.

[23] M. F. Santos, “Sistemas de Classificação em Ambientes Distribuídos”, Ph.D. thesis, Universidade do Minho (in Portuguese), 1999.

[24] A.Abelha, J. Machado, M. F. Santos, S. Allegro, F. Rua, M. Paiva and J. Neves, “Agency for Integration, Diffusion and Archive of Medical Information”, presented at the IASTED International Conference on Artificial Intelligence and Applications, Benalmádena, Spain, September 2003.

[25] P. Gago and M. F. Santos, “Evaluating hybrid ensembles for Intelligent Decision Support for Intensive Care”, In O. Okun, and G. Valentini, (Eds.) Supervised and Unsupervised Applications of Supervised and Unsupervised Ensemble Methods, Vol. 245, pp. 251-265, Springer, 2009.

Automation Techniques of Building Custom Firmwares for Managed and Monitored Multimedia Embedded Systems

J. Slachta, J. Rozhon, F. Rezac, M. Voznak

Abstract**—**One of the biggest challenges facing network administrators is the management of increasing amount of devices that are under their administration. The article deals with solution how to automatically build a system image with communication server and with advanced techniques of automated configuring such devices based on OpenWrt Linux distribution. The solution is built as a universal open source modular system and the server has been developing within the framework of a BESIP project (Bright Embedded Solution for IP Telephony) since May 2011. This open-source modular system with overall concept and the architecture is described in detail in this paper.

Keywords**—**Build system, Provisioning, OpenWrt, VoIP Security, SIP.

I. Introduction

THE aim of the BESIP project is the development and implementation of embedded SIP communication server. Among all desired characteristics mainly belongs an easy integration into the computer network based on open-source solutions. This project serves as a secure and robust SIP IP telephony infrastructure available for anybody. It offers the prepared solution with integrated key components and the entire system is distributed as a firmware image or individual packages that might be installable from repositories. The main goal is to provide a solution which should be easily installable and configurable even without the deep knowledge of the technologies that are used by our key components. Also it aims to be scalable solution with unified configuration in mind [1].

Several open-source applications were adopted and implemented into developed modules, however within the implementation many modifications were required, especially in the core module (OpenWrt) due to complicated porting of applications into OpenWrt Buildroot. Our patches were verified and accepted by OpenWrt community. The speech quality monitoring tool was developed from scratch and implemented in Java. BESIP can run on embedded devices as well as on high performance devices. It requires at least 32 MB RAM and runs on the majority of OpenWrt supported devices [2],[3].

II. State of the Art

As mentioned in the introduction, we discuss the implementation of a SIP communication server solution which would be an alternative to several current implementations. The main advantage of our solution is the ability to easily and quickly set up a full featured PBX on almost any hardware. We can presume that almost all implementations are based on open-source Asterisk PBX, web-interface for Asterisk and with a GNU/Linux distribution on the base layer.

At present, there are several projects that offer multipurpose IP telephony solutions for embedded devices and for household or enterprise platforms. The initial project of a GNU/Linux distribution which offers an easy set-up of IP telephony in a few steps is the Asterisk@Home project. The first version of this project was released on 29 April 2005. This project integrated a web interface for Asterisk, Flash Operators Panel to control and monitor PBX in real-time and also offered a full FAX support within one bootable image for almost any x86 PC. On 3rd May 2006 the development of this project was discontinued and was replaced by its successor Trixbox. However, the development of Trixbox does not seem to continue any more. Two existing projects - AsteriskNOW and Elastix – now offer an alternative to Trixbox.

The former, AsteriskNOW appears to be similar to Trixbox – a packed GNU/Linux distribution with Asterisk with a FreePBX web interface on top of it.

The latter, Elastix, is a bit more modular. Compared to any other project, it offers a slightly more modular hierarchy to facilitate the applicability to a multiple service server. The

This research has been supported by the Ministry of Education of the Czech Republic within the project LM2010005..

J. Slachta is a M.S. student with Dept. of Telecommunications, Technical University of Ostrava and he is also a researcher with Dept. of Multimedia in CESNET, Zikova 4, 160 00 Prague 6, Czech Republic (e-mail: slachta@cesnet.cz)

J. Rozhon is a PhD. student with Dept. of Telecommunications, Technical University of Ostrava and he is also a researcher with Dept. of Multimedia in CESNET, Zikova 4, 160 00 Prague 6, Czech Republic (e-mail: rozhon@cesnet.cz).

F. Rezac is a PhD. student with Dept. of Telecommunications, Technical University of Ostrava and he is also a researcher with Dept. of Multimedia in CESNET, Zikova 4, 160 00 Prague 6, Czech Republic (e-mail: filip@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.

.

increasing popularity of embedded devices, such as Raspberry Pi, is the reason why the Micro Elastix distribution was born. However, all of those projects are either prepared for x86 machines only or for specific hardware. Micro Elastix only supports three platforms, namely PICO-SAM9G45, Raspberry Pi and MCUZONE.

None of the projects includes a security module that would offer a complete IPS and IDS system to prevent attacks against the SIP Registrar server. Also there is no module that would monitor the quality of voice calls transmitted through an integrated PBX. Thanks to the portability of the OpenWrt distribution we can prepare a BESIP bootable image for almost any device.

III. Architecture of BESIP

One of the biggest challenges during BESIP development was to create or modify any existing Linux distribution to serve our expectations. We needed to create an environment that would be fully customizable to any purpose and also to be easily maintainable through the time the BESIP would be developed. The advantage of portability to any platform was also welcomed. The choice of Linux distribution we wanted to modify fell on OpenWrt Linux distribution. The reason why we chose that system was the approach for building firmware, the toolchain, crosscompiler and all applications are downloaded, patched and built by scratch. This means that OpenWrt does not contain any source code, it does only have its build system with templates, patches and Makefiles with procedures how to build a system and its packages for targeted device. This approach allows us to create custom procedures for build system and packages that can be modified at any stage.

A simplified view on BESIP architecture is depicted in Fig. 1 which describes how the architecture is designed. The first block, the build system, is a wrapper on the top of the OpenWrt build system. It is designed for easy creation of firmware images within the single text file which describes what should be built for specific architecture and device we are targeting on. With the build system comes also several BESIP packages that are customizable from the OpenWrt buildroot, e.g. before the firmware is built.BESIP packages consists of several modules which provides functionality as:

- the PBX module to accomplish VoIP functionality,
- the Monitoring module to monitor speech quality and the system itself,
- the Security module to provide IPS/IDS system,
- the Core module as a glue to all services among themselves and to provide intuitive interface to them.

The security module is based on SNORT; SNORTSam and iptables [4]. In addition to this, the Kamailio ratelimit and pike module is used for defending attacks.

The monitoring module exploits a tshark package and our java code which interprets its results and gives information about particular speech quality. The Zabbix agent is used to report basic states of the entire system and finally the PBX module is made from Kamailio in conjunction with Asterisk.

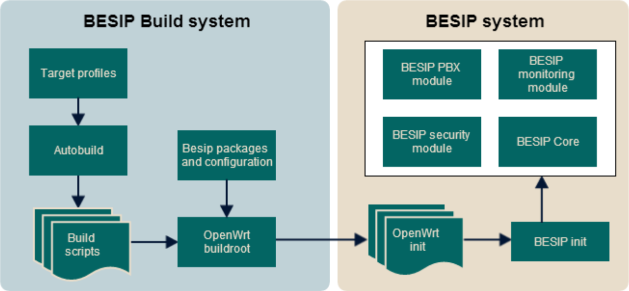


Fig. 1 Architecture of BESIP system.

The Core module is a shell library (providing functions for all executable BESIP scripts) with executable files which makes all mentioned services fully working.

IV. Build System for Besip

Before we describe the concept of BESIP system, it is necessary to introduce the BESIP build system which makes automation of creation system images much easier. As said above, BESIP is based on GNU/Linux distribution OpenWrt which is built on top of the OpenWrt Buildroot. Buildroot is a set of Makefiles and files that allows to compile cross-compilation toolchain and to generate by that toolchain resulting cross-compiled applications into a root filesystem image to be used in a targeted device. Cross-compilation toolchain is compiled by host compilation system which is provided by any GNU/Linux distribution.

In the beginning of BESIP development, we met issues that were holding us back. We could not test all changes immediately, we had to recompile all code and generate images nearly always when we ported new application, modified post installation scripts or when cross-compilation toolchain has changed. Also, the system behaves differently during testing if it is new root filesystem image, or modified root filesystem that has been run more than once. At least those issues led us to create an easy interface that will ease the creation, automation and functional testing for system images.

BESIP build system is a set of scripts, Makefiles and definition files that make an easy interface to OpenWrt Buildroot. We can consider the main Makefile to be as a core of the BESIP build system. It performs all atomic operations with OpenWrt Buildroot, works with source code management systems (to update/revert/any operation with local copies of OpenWrt source codes), patches OpenWrt Buildroot and executes images as virtual machines. Those commands might be used by any user or by autobuild scripts, which will be described after.

On the top of the core Makefile is autobuild script. This script calls all atomic operations within more complex parameterized operations whose variables are defined in specific target files. Those target files are user defined and on the basis of those files are configuration files for OpenWrt

Buildroot created. Once we have configuration files the system images could be created by calling autobuild.sh script with command *build* and parameter containing the name of the target file.

The following simplified example shows, how to create target file.

```
TARGET_CPU=x86
OWRT_NAME=trunk
TARGET_NAME=virtual_\$(BESIP_VERSION)-
owrt_\$(OWRT_NAME)
TARGET_QEMU=i386
TARGET_QEMU_OPTS=-m 512
OWRT_IMG_DISK_NAME=openwrt-\$(TARGET_CPU)-
generic-combined-squashfs.img
BESIP_PACKAGES= gnugk=y suricata=y
EMBEDDED_MODULES += SATA_AHCI VMXNET3
```

The following example shows how to build system images based on the target file.

```
#./autobuild.sh build virtual-x86-trunk
```

Such techniques can be used for any purpose of automated building system images for any device or platform supported by OpenWrt. Those could be firmware images for campus access points, specialized network probes, virtualized multimedia servers or any other devices.

V. Core Module

The role of the Core module is to provide a glue among all services that served by all BESIP modules. The most important part of the Core module is the BESIP shell library that provides functions for all utilities and scripts used by BESIP system. Functionality of a Core module complements utilities for configuration management and for simplified configuration of system image. With all those utilities comes along also default configuration which prepares all module services into fully functional state with all BESIP modules running and operational. Also, the role of this module is to switch any existing OpenWrt environment to BESIP environment while the device is booted the first time or the BESIP environment is used and ran the first time.

A. BESIP Environment

BESIP environment handles tasks which has to be performed at several stages in OpenWrt operating system. After the BESIP firmware image is built at this stage the operating system behaves as clean operating system with installed dependencies required by BESIP package and its submodules. On the first boot the init script is performed and it waits until the overlay filesystem is mounted. When the filesystem is mounted the *first_boot* procedure is performed. This procedure incorporates the initial setup of the system and preparation configuration files. The main advantage of this procedure is the applicability to any existing setup of OpenWrt system.

Another mandatory part of BESIP system is an executable application that provides functions to manage following procedures:

- Generates provisioning data for connected phones,
- resetting system image into factory defaults,
- performs system upgrade,
- controls internal BESIP modules,
  - o configuration and management of security module,
  - o importing and setting up a dial plan for PBX module,
- collects information for crash reporting to be used for debugging.

B. Provisioning Client

The impetus for development of provisioning tool arose during the period when firmware images created by BESIP build system were deployed to computers, routers and wireless access points. Those machines were not configured for target networks, which were supposed to be deployed on. Because the target configuration does not depend on a person which builds the system, but on the network administrator, then configuration should lay outside of a BESIP firmware image. The creation of such tool bring a question how should the target device should fetch and apply its configuration.

In the build system, we can pass static information about our provisioning server which provides configuration (during build time). We can also change this information in firmware image. This information can be used for protocols which translates one kind of information to another. As an example we can use DNS protocol and its TXT records. The target configuration could be stored on a server designated within an URI in a variable from TXT record which is obtained from static URL provided by BESIP build system. This solution is replicable for any protocol which allows distribution that kind of information (LLDP, DHCP or any other else).

An example how to resolve UCI provisioning URI:

```
host -t txt provdomain
provdomain descriptive text
"provuri=http://12.34.56.78/uciprov/"
```

If a device knows where to obtain configuration from then the device can construct all provisioning URI addresses for each device state it needs. This approach is needed when system administrator needs to differentiate configuration for devices which starts up the first time, if those devices are refreshing its common device configuration on a regular basis or if it is the configuration that is obtained after device startup. UCI provisioning client written for BESIP currently handles only configuration files that are handled by UCI system (Unified Configuration Interface) for centralized configuration. If a device knows where to obtain configuration from, then the device can obtain configuration data from ordinary transport protocols designated in provisioning URI. The benefits that BESIP draws from OpenWrt builds upon the UCI configuration system which is based on plain text configuration files with firmly defined structure. This configuration is obtained using software for file retrieval from network resources, e.g. wget, and immediately imported into UCI.

The client side of UCI provisioning currently has several stages:

1. Waiting for the system to be ready to be provisioned,
2. Stage 1 - receive provisioning URI using supported protocols,
   - For each supported protocol try to receive provisioning URI address,
   - Construct a list of URI provisioning addresses based on received URI address.
3. Stage 2 - obtain configuration from URI received in stage 1,
   - For each interface try to obtain configuration data.
4. Stage 3 - apply received configuration.

The server side of UCI provisioning is currently solved by providing static file structure with files which consists of export provided by UCI system. See sequence diagram depicted in Fig. 2 to see how the UCI provisioning works.

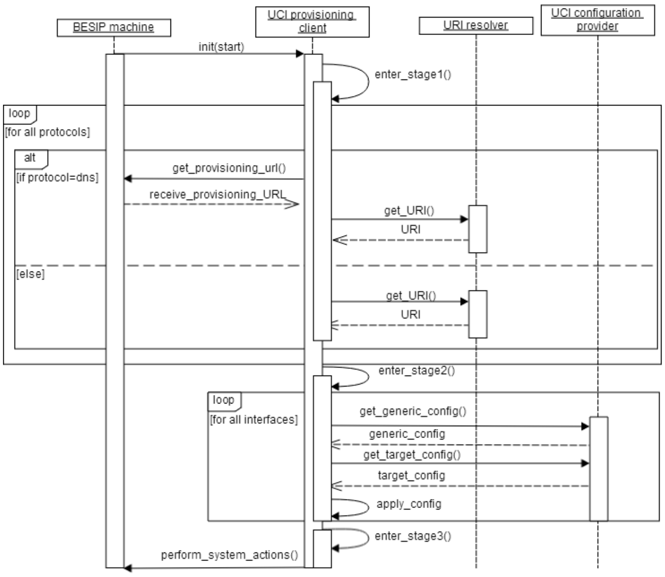


Fig. 2 Sequence diagram of UCI provisioning client.

VI. PBX Module

The PBX module is a key part of the BESIP project. It operates as SIP proxy or SIP B2BUA, depending on configuration, and ensures a call routing. Asterisk is used for call manipulation and for the PBX functions. Kamailio is used for the proxying SIP requests, the traffic normalization and for the security [5]. There are always two factors when developing VoIP solution the first one is high availability and reliability, the second one is an issue of advanced functions. Many developers try to find a compromise, we have implemented both, and our BESIP is able to adapt to the users requirements. More complex system can handle many PBX functions such as a call recording or an interactive voice response but due to the bigger complexity it is more susceptible to fault. On the opposite side, pure SIP proxy is easier software, which can perform call routing, more fault tolerant, but it is more difficult to use the advanced PBX functions [6].

VII. Security Module

Security module is very important part of BESIP and all the time, it was considered to make the developed system as secure as possible. Next to this, entire system has to be fault-tolerant, monitored and protected from attacks. It means that if the device is under attack, only attacker has to be blocked, not entire system or other users. If there is some security incident, BESIP immediately solves the situation and notifies this event in a detailed report to administrator.

The attack are recognized and processed by SNORT rules, the source IP address is automatically sent into the firewall by SNORTSam and the intruder’s IP is blocked. This is very flexible, reliable and effective implementation. Dropping attack based on IP directly in the Linux kernel is much more efficient than to check messages on the application level. Only first messages are going to SNORT filter. When SNORT identifies a suspicious traffic, next messages from the same IP are blocked.

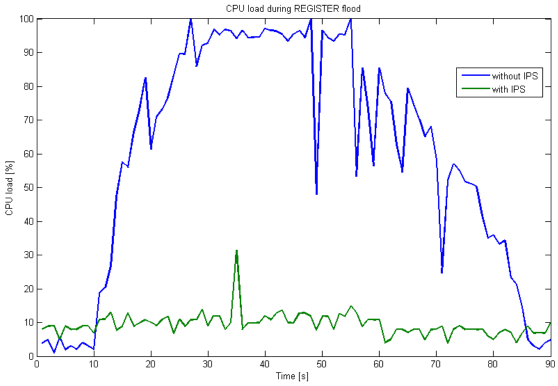


Fig. 3 Attack effectiveness based on REGISTER flood.

If more soft faults appear from some IP, it is blocked at the IPTABLES level; this approach can effectively block incorrectly configured clients and servers. For example, if a client sends REGISTER with proper credentials, it is not obviously security attack but the client attempt to register again and again, with every registration requires computing sources at SIP REGISTRAR server. Such attempts can be denoted and blocked for a time interval. Administrators can use Zabbix agent inside BESIP to gather all information directly into their monitoring system. The monitoring is very important part of the security module and BESIP team was already focused on the issue in early design [7]. Partially, BESIP is resistant to some kind of DoS attacks. It depends on hardware used. If the hardware is strong enough to detect some security incidents on application level, the source IP is immediately dropped. Low performance hardware cannot handle such detection on application level. In such case, it is better stop DoS attacks before it reaches BESIP. For example, SNORT on a dedicated machine will be much more flexible than if is an integral part of VoIP system. Therefore, we recommend using an external IPS system to make VoIP service robust and secure. Nevertheless BESIP includes own IPS/IDS system [8], [9].

The features of our security module were verified in test-bed and results are depicted in Fig. 3 [7]. The CPU load was monitored during trivial SIP attacks. The line SSI (Snort, SnortSam, IPtables) represents the response in case of active security module in BESIP whereas next dependencies were measured without SSI. There were emulated only two types of DoS attacks, namely REGISTER flood and INVITE flood. In order to generate these attacks, we used sipp generator and in case of INVITE also inviteflood tool. The dependencies in both figures clearly prove the ability of security module to mitigate the performed attacks.

VIII. Conclusion

As we have mentioned, BESIP consists of several components, which are distributed under GPL as an open-source solution. A few of them have been fully adopted such as components in Security and PBX modules, some of them modified, concerning the Core module and finally we have developed own tool for Speech quality assessment. The contribution of our work is not only hundreds of hours spent on the development, on the coding BESIP system, we bring a new idea of the unified configuration management, with unified CLI syntax which enables to configure different systems, Asterisk and Kamailio in our case.

BESIP is distributed as a functional image for several platforms, mainly for x86 platform, which is also possible to run it on any virtualization x86 software. There are several example firmware images for several target devices, such as TP-Link access points or Raspberry PI computer. Configuration is available through web-browser, SSH client or to be provisioned using supported provisioning protocols. After the testing, version 2.0 will be released; a new release 2.0 will be based completely on NETCONF with one API to configure the entire system. Next to this, CLI syntax has been developing and will be connected to NETCONF. CLI will be independent of internal software so if some internal software is modified, there will be no change in configuration. Even more, CLI and NETCONF configuration will be independent on hardware and version. To export configuration from one box and to import it to the next one will be a simple task. Users will modify only one configuration file to manage entire box. Project pages are available at [10], binary images from the auto-build system can be downloaded from [11] and source codes can be checked out via SVN from the same page as well [11].

References

[1] M. Voznak, F. Rezac: “Threats to voice over IP communications systems”. *WSEAS Transactions on Computers*, Volume 9, Issue 11, 2010, pp. 1348-1358.

[2] F. Abid, N. Izeboudjen, M. Bakiri, S. Titri, F. Louiz, D. Lazib: “Embedded implementation of an IP-PBX/VoIP gateway”. *24th International Conference on Microelectronics*, December 2012, IEEE, Article number 6471377.

[3] N. Titri, F. Louiz, M. Bakiri, F. Abid, D. Lazib, L. Rekab: “Opencores /Open-source Based Embedded System-on-Chip Platform for Voice over Internet”. *INTECH: VOIP Technologies*, pp. 145-172.

[4] J. Safarik, F. Rezac, M. Voznak: “Monitoring of Malicious Traffic in IP Telephony Infrastructure”. *Technical Report*, 10p., December 2012.

[5] M. Voznak, J. Safarik: “DoS attacks targeting SIP server and improvements of robustness”. *International Journal of Mathematics and Computers in Simulation*, Volume 6, Issue 1, 2012, pp. 177-184.

[6] J. K. Prasad, B. A. Kumar: “Analysis of SIP and realization of advanced IP-PBX features”. *3rd International Conference on Electronics Computer Technology*, Volume 6, 2011, IEEE Article number 5942085, pp. 218-222.

[7] M. Voznak, K. Tomala, J. Vychodil, J. Slachta: “Advanced concept of voice communication server on embedded platform”. *Przeglad Elektrotechniczny*, Volume 89, Issue 2 B, 2013, pp. 228-233.

[8] D. Endler, M. Collier: “Hacking Exposed VoIP”. McGraw-Hill Osborne Media, 2009.

[9] D. Sisalem, J. Kuthan, T.S. Elhert, F. Fraunhofer: “Denial of Service Attacks Targeting SIP VoIP Infrastructure: Attack Scenarios and Prevention Mechanisms”. IEEE Network, 2006.

[10] Management of BESIP Project, LipTel Team, 2014, https://besip.cesnet.cz

[11] Project BESIP, https://homeproj.cesnet.cz/projects/besip/wiki/Download

Mobile Augmented Reality and Location Based Service

Sagaya Aurelia, Dr. M. Durai Raj, Omer Saleh

Abstract-- Mobile Augmented Reality(MAR) is characterized as a technology providing the same feature as Augmented Reality(AR), but without the physical restrictions of a research facility or a testing area location. A Location-Based Service (LBS) is a mobile computing application that provides services to users based on their geographical location. In the course of the rise of mobile devices with more and more functionalities (especially Apple's iPhone and Android-based devices), location-based services constantly grow in popularity. More and more information is enriched with geodata and thus can not only be presented in a virtual space, but in real, mobile contexts and in a context-sensitive way adapted to the user's preferences. This paper states analyzes the concepts and advantages of mobile augmented reality and location based service and the combination of mobile augmented reality along with location based services. The challenges along with pros and cons are discussed.

Keywords-- GIS, GPS, human computer interaction,location based service, Mobile augmented reality

I. Introduction

Augmented reality (AR) is a field that intertwines various topical technologies and emerging concepts. Mobile Augmented Reality(MAR) is characterized as a technology providing the same feature as Augmented Reality(AR), but without the physical restrictions of a research facility or a testing area location. Mobile AR utilizes various sensors to create a picture of the surroundings and to infer what digital content relates to the current context. There are several tracking methods, varying from large-scale solutions based on GPS, GSM or wireless-LAN to more accurate ones based on magnetic fields, intertial solutions (accelerometers and gyroscopes), sensors (e.g., radio frequency identification), visual markers or markerless tracking. For example, GPS is useful for aligning the AR content over long-distances, but often too inaccurate inshort distances (<50m) (Thomas et al. 2002. The different solutions can more or less be used in any kind of environments but there is variation in the achieved accuracy and range of use. With the help of such technologies, the AR system can infer, for example, the user’s location, what she is looking at, and to where and how fast she is moving [13].

A Location-Based Service (LBS) is a mobile computing application that provides services to users based on their geographical location.

Consequently, this paper presents MAR from various viewpoints, starting with a description of the concept of MAR in general, MAR content, advantages. It is followed by definition, architecture, location based service processing and applications of Location based service. The combination of Mobile augmented reality and Location based services is discussed further. Finally concluded along with the challenges [13]

II. Augmented Reality

Augmented reality (AR) is a field that intertwines various topical technologies and emerging concepts. AR is a multifaceted term that can refer to (1) a technology or a group of technologies – a mesh-technology that utilizes also several other technologies, (2) a concept that describes a vision of future computing, (3) a field of research in various disciplines, (4) a medium and an interface to digital information, and (5) recently also a platform for creating novel services and business. The reality virutality continuum is shown in figure 1.

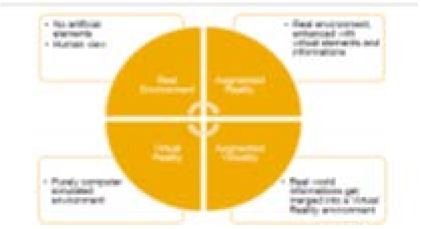

No artificial elements
Human view
Real environment enhanced with virtual elements and informations
Real Environment
Augmented Reality
Virtual Reality
Augmented Virtuality
Purely computer simulated environment
Real world informations get merged into a virtual Reality environment

Fig. 1 The reality- Virtuality Continuum

III. Mobile Augmented Reality

As a result of the rapid advancement of mobile devices, AR is entering also the mobile domain. After this, smart phones have been equipped with integrated cameras, sensor technologies like GPS and orientation sensors, high-resolution full color displays, highspeed networking, high computing power, dedicated 3D graphics chips etc. as shown in figure 2. For example with regard to the sensor technology, smart phones can serve as external eyes and ears for sensing embedded information in the surrounding environment. Such a plethora of possibilities being integrated in one device that is extensively spread provide a dexterous platform for building AR applications and services (Wagner & Schmalstieg 2009, Henrysson 2007) [13].

Mrs P. Sagaya Aurelia is with Barathidasan university as part time research scholar, presently working in Azzaytuna university, Libya (e-mail: psagaya.aurelia@gmail.com).

Dr. Durai Raj is with Department of Computer Science, Bharathidasan university, Trichrapalli, India(phone no: 0919487542202,email:durairaj.bdu@gmail.com)

Dr Omer Saleh is currently the Director of Post graduate, research and Development cum Head of the Department, Department of computer science, Faculty of Education, Azzaytuna University, Libya(phone no: 0926895760,e-mail:Immer.jomah@gmail.com)

However, mobile AR is not only about having a mobile as shown in figure 2 or hand-held device as hardware. It is about AR being enabled for truly mobile and ubiquitous contexts and activities – instead of the use being tied to stationary locations and carefully conditioned environments, such as in medical or manufacturing applications of AR (Höllerer & Feiner 2004). Mobile contexts and activities with AR could include, e.g., information search 'in the wild', wayfinding, choice of services and products, social interaction, entertainment and exploration of larger areas. Additionally, for example military applications and maintenance could utilize both mobile and stationary AR [13].

High performance processors
Powerful graphics
Wireless networking
Multimedia
Compass direction
Camera
Orientation
GPS Location

Fig. 2 Everything together in mobile phone [1]

Feiner and Hollerer[18] identified six components necessary to provide true MAR:

- Computational platform to process all relevant information, and to compute the visualization of AR objects presented on the display.
- Display to present the virtual objects to the user.
- Registration of environment. Registration of camera input and head orientation helps to present the AR objects correctly aligned with the real world.
- Wearable input and interaction technologies to enable a mobile person to work and collaborate with other users
- Wireless networking for instant communication with other people and central databases
- Data storage and access technology to provide the user with all context relevant data in the environment intended for augmentation.

IV.CONTENT OF MAR

The digital information content shown inMARcan relate to and augment anything in the user's current context, for example physical structures, places, things in nature, moving things like products and people, and also intangible and abstract things like services and events. The content varies in form, consisting of, for example, 3D multimedia content, 2D graphics, animated graphics, frames highlighting objects or their shapes (e.g., corners, planes) in the reality, simple textual information and graphical symbols. Wither et al. (2009) refer to content as annotations: "additions of extra virtual information to an object". The various annotations can, for example, simply provide a name of the object in reality, describe its characteristics (e.g. availability of a service), add new virtual objects to the scene (e.g. virtual characters), modify the real objects (e.g., change surface colour or luminosity), or direct the user with arrow*s and other highlights (Wither et al. 2009). In other words, AR content can be added either directly about a particular real world object or shown in a more indirect or abstract way. Furthermore, the semantic relevance and permanence of the annotation depends on the user's task, her interactions, and changes in the virtual elements of AR.

The origin and storage place of the content can naturally be the local device (e.g., 3D models stored in device memory) or, currently more common, various online repositories to access with a network connection. MAR is a fruitful interface for exposing large amounts of visual content from existing online services like Wikipedia and content sharing services like Flickr[15]. With the development and openness of "network societies" (Castells 2000), for example map and multimedia repositories are being democratized by public authorities

A large portion of this Internet-based content is geo-tagged or otherwise bound to a location, which has made location-based MAR a rapidly growing area. Especially user-created geotagged content has become increasingly common thanks to online maps with user-created point-of-interest (POI) information, and online services built around maps (e.g. Yelp2 [15]). The location-based content can be efficiently related to the real world – both technically and mentally. The location information ties the situation in a certain space and helps delimiting what content to show on the AR interface [13].

V.ADVANTAGES OF HANDHELP VIDEO-SEE THROUGH MAR:

Compared to desktop or other mobile technology use, the interaction moves towards context-based use where the service of the technology depends on the surrounding information and the user's activities.

1. MAR can be seen as a local search engine to the information embedded in the environment.

2. MAR provides a tangible interaction metaphor for utilizing the realms of digital information.

3. MAR is a lens-based UI: it provides an inherently limited display size as well as a limited field of view (window to the AR, Milgram & Kishino 1994) [13].

VI. LOCATION BASED SERVICE

GPS and cellular positioning are both used today in mobile devices, such as the iPhone. There are many applications which use these two technologies together.

LBSs contain a number of components including maps and Geographic Information System (GIS) information, location collection services, and LBS application-specific subcomponents. The architecture of an LBS can be generalised as shown in Figure 3. [2]

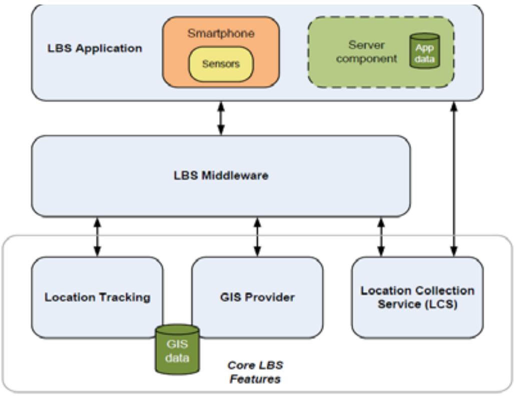


Fig. 3 Components of LBS [2]

Component	Description
LBS Application	This represents a specific application such as a “find my friends” application. This consists of a smartphone component, which has a number of sensors, and potentially a server component that includes application-specific data
LBS Middleware	This wraps access to Core LBS Features (Location Tracking, GIS Provider and Location Collection Services) to provide a consistent interface to LBS applications. The OpenLS specification represents one standard for LBS middleware
Location Tracking	This component stores the location trace of individual users. This represents a fundamental component in next-generation LBS as it contains the data that allows a u potentially predicted. In particular, this component would typically support. Keep records on user s̋ currentlocations . and pas. Notify other components when a specific user has moved, or based notifications being sent to users. Determine which users are within a defined location. This supports geocasting features. Queries of location trace to generate user movement models.
GIS Provider	This component provides geospatial functionality for many LBSs including map information, map visualisation and directory services Google Maps with its API can be considered a GIS provider
Location Collection Service (LCS)	This component performs location collection to get a latitude and longitude for a specific user. Depending on the technology, this component may be accessed via the LBS Middleware (e.g., mobile network triangulation via a service provider) or directly (e.g., via GPS receiver in the smartphone).

Table 1. Description of Components of LBS[2]

VII. LBS COMMUNICATION MODEL:

The LBS communication model consists of three layers – a positioning layer, a middleware layer, and an application layer (Schiller & Voisard, 2004) (see Figure 4). The positioning layer is responsible for calculating the position of a mobile device with the help of a position determination equipment and the geospatial data in a geographic information system. The calculated position is then passed directly to an application. Recently mobile network operators have introduced a middleware layer between the positioning layer and the application layer to reduce the complexity of service integration, saving operators and third-party application providers’ time and cost for application integration. This middleware layer manages the interoperability between networks for location data [3].

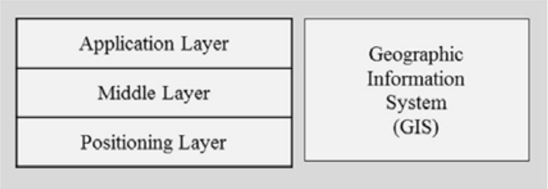


Fig. 4 The LBS Communication Model [3]

The Summary of GPS and Wi-Fi based location collection technologies are shown in table 2 and table3.

- GPS-based solutions

	GPS	Assisted GPS
Description	The device’s postriagulated based on signalsfrom at least four GPS satellitesbased on the known position ofthe	This is an enhanced form of GPScommonly usedon smart-phones, in which an “assistance”server on themobile network provides

	satellites, the time thatmessages from the satellites weresent and the time that they werereceived.	informationsuch as accurate GPS satellite orbitinformation, accurate timestamps orpossibly snapshots of GPS signals. Thiscan allow GPS accuracy with initiallocation information within seconds,thereby making it practical for use inLBSs.
Accuracy	5-10 mHighly accurate. No dependencyon a mobile network provider.	5-10 mHighly accurate,and allows GPS to beused in more areas, such as in densely. Populated areas where clear GPSsignals may not be obtainable. Fastlocation collection.
Cons	Relatively high powerrequirement, as a GPS receiverneeds to operate.It can only be used outdoorswhere clear satellite signals canbe obtained.6Depending on the device, it maytake a long time (~30 seconds) tolock onto satellite signals.	There is a dependency on the mobilenetwork provider; it can only operatewhere mobile network reception isavailable.

Table 2. Summary of GPS-based location collection technologies [2]

- Wi-Fi-based solution

	Wi-Fi Positioning System
Description	The identities and relative signal strengths (which correspond roughly todistance) to public Wi-Fi access points are recorded by the device, therebyallowing triangulation with respect to these access points. Based on adatabase of known access points and their physical location, an approximatelocation for the device can be calculated. This system was initially created bySkyhook.7
Accuracy	Fast and relatively accurate location collection compared to mobile networktechniques. Lower power requirements compared to GPS due to speed and no need for GPS receiver. Allows devices such as laptops to use location collection and hence interact with some LBS applications (such as those equipped with HTML 5 Geolocation8).
Cons	Relies on access to Wi-Fi access points, which may not be available in certainlocations.

Table3. Summary of Wi-Fi-based location collection technologies [2]

VIII. Information for searching, identifying and checking

The two basic actions locating and navigating mainly rely on geospatial information. Searching, identifying and checking however need a bigger variety of different information. Additionally to the geospatial information also other types of information are needed:

Comprehensive static information are mainly contents such as a yellow pages. Such information stays constant over a while and could of course also be retrieved via other media (book, newspaper, map, TV, internet, etc.).

Topical information that may change while the user is on the move. In such a case the information checked previously from other media may no longer be valid. Examples of such topical information are traffic information, weather forecasts, last-minute theatre ticket deals, or on-line chat.

In addition to topical information, the users will need guidance on how to proceed in the changed situation. For instance, a train schedule as such can be obtained elsewhere but once on the move, the user will need information on delays and estimated arrival times.

Additionally safety information has key importance, e.g. actual information on the state of the roads or hiking trials, weather changes, danger of falling rocks, etc. Car drivers or boaters also need information in emergency situations, e.g. roadside help in a situation when the car breaks down.

Far too often users are seen as passive information consumers

	Action	Questions	Operations
	orientation & localization locating navigation	Where am I?Where is {person|object}?	positioning, geocoding,geodecoding

	navigating through space, planning a route	How do Iget to placename| address| xy}?	positioning, geocoding,geodecodingrouting
	search searching for people and objects	Where is the {nearest | most relevant | &}{person| object}?	positioning, geocoding,calculating distance andarea, finding relationships
	Identification identifying and recognizing persons or objects	{What | who | how much} is {here |there}?	directory, selection, thematic/ spatial, search
	event check checking for events; determining the state ofobjects	What happens {here |there}?	

Table 3 User activities Reichenbacher (2004).[4][5]

IX. CATEGORIES OF LOCATION SERVICE APPLICATIONS

There exist a broad range of different location based services. The figure 5 gives an over-view on the main categories of LBS applications. This listing does not claim to be complete and is certainly growing over time. For some application fields, namely navigation, information, advertising & billing and games & leisure, additionally information on the positional accuracy needs, the environment and the service type (push or pull service) are shown as graphics in figure 6.

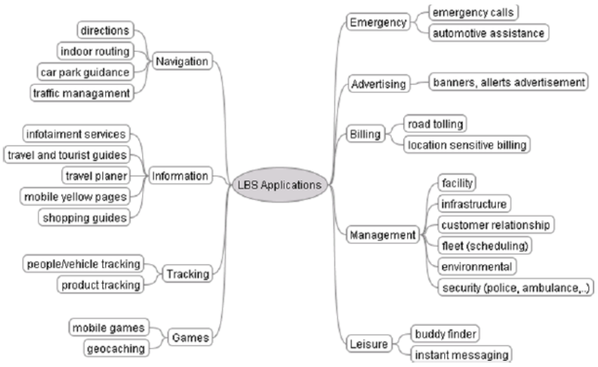


Fig. 5 Overview of LBS application [4]

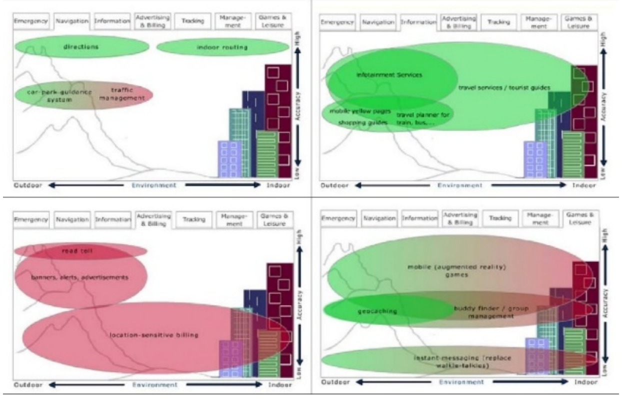


Fig. 6 Properties of a selection of LBS applications High positional accuracy denotes an accuracy within 50 meter while a low accuracy is worse than 300 meter. Red: push service, Green: pull service [4]

X. LBS SERVICE REQUEST PROCESSING

Considering the example of searching an Indian restaurant the information chain from a service request to the answer will be described in the following and is illustrated in Figure 7 information the user

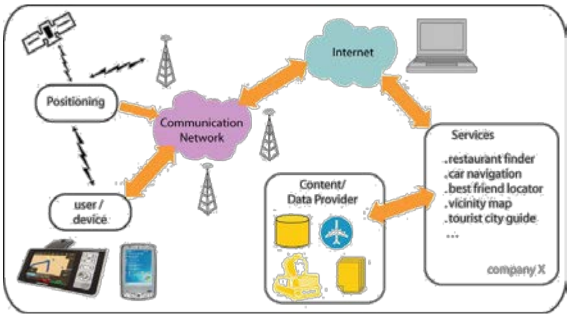


Figure 7. LBS components and information flow [4].

The want is a route to an Indian restaurant nearby. Therefore the user expresses his need by selecting the appropriate function on his mobile device: e.g. menu: position information => searches => restaurants => Indian restaurant.

Now if the function has been activated, the actual position of mobile device is obtained from the Positioning Service. This can be done either by the device itself using GPS or a network positioning service. Afterwards the mobile client sends the information request, which contains the objective to search for and the position via the communication network to a so called gateway.

The gateway has the task to exchange messages among mobile communication network and the internet. Therefore he knows web addresses from several application servers and routes the request to such a specific server. The gateway will store also information about the mobile device which has asked for the information.

The application server reads the request and activates the appropriate service

Now, the service analyses again the message and decides which additional information apart from the search criteria (restaurant + Indian) and user position is needed to answer on the request. In our case the service will find that he needs information on restaurants from the yellow pages of a specific region and will therefore ask for a data provider for such data.

Further the service will find that information on roads and ways is needed to check if the restaurant is reachable (e.g. sometimes a restaurant on the other river side might not be reachable since no bridge is nearby).

Having now all the Information the service will do a spatial buffer and a routing query (like we know from GIS) to get some an Indian restaurants. After calculating a list of close by restaurants the result is sent back to the user via internet, gateway and mobile network.

The restaurants will now be presented to the user either as a text list (ordered by distance) or drawn in a map. Afterwards the user could ask for more information on the restaurants (e.g. the menu and prices), which activates a different kind of services. Finally if he has chosen a specific restaurant he can ask for a route to that restaurant[4].

XI. MAR AND LBS

A vital part of Augmented Reality (AR) is to create a credible experience. In AR, objects are superimposed on a real world. To create a convincing superimposed object, the object need to be aligned with the surroundings. To achieve this the user's location and orientation needs to be accurately tracked. When tracking the current position and subsequent movement of the user, the application can use different methods to retrieve up-to-date position information [4].

The available tracking methods when applying AR in an application on an iPad device, depends on the connectivity features in the iPad. In an unaltered iPad, tracking of the user can be conducted by two methods. The method applied can be a static identicator or a dynamic identicator.

A special kind of such location-based services are augmented reality services that provide a computer-supported, extended reality by displaying relevant information in the user's environment. With the new generation of mobile devices and available “reality browsers“, there is for the first time an infrastructure that allows for the creation of augmented reality services without the need of a complex instrumentation and the development of respective interfaces. Thus, the plenitude of localized information can principally be made available to end users in different scenarios by means of augmented reality services, depending on the users' locations as well as their preferences and contexts.

For a broader picture, Figure 8 reflects AR to specific well-known interface types that can be seen as prior interfaces for accessing and browsing digital information. Map interfaces refer to 2D representations of physical areas from above (possibly with POIs), 2D AR refers to ego-centered AR where tracking is based on GPS and magnetometers and simple 2D augmentations are visualized, and 3D AR refers to augmentations being 3D and precisely aligned. The figure summarizes the physical scale of interaction in relation to each of the interfaces and how various aspects change along this continuum. When moving from traditional WWW towards 3D AR, the integration of realities and reproduction fidelity grow, and the precision of the augmentations increases. Accordingly, there is an increase in the contextuality and relevance of the accessed information content and the user's sense of presence in the reality that the interface displays [13].

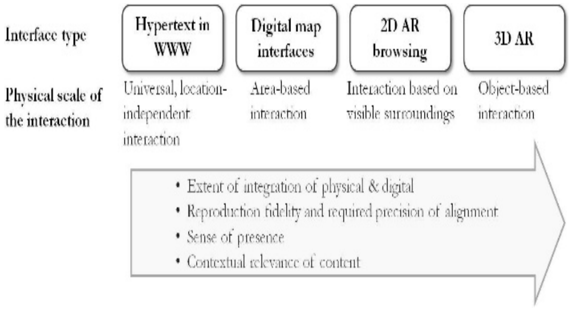

Interface type
Hypertext in WWW
Digital map interfaces
2D AR browsing
3D AR
Physical scale of the interaction
Universal, location-independent interaction
Area-based interaction
Interaction based on visible surroundings
Object-based interaction
• Extent of integration of physical & digital
• Reproduction fidelity and required precision of alignment
• Sense of presence
• Contextual relevance of content

Figure 8. AR in a continuum of interfaces for accessing digital information [13].

A. Using the location

When the mobile device has a set of coordinates to use, they can be sent to a server in a request for data associated with them, such a nearby planning applications. These can either be displayed in AR through the users camera view, or placed on to a map to take the user to the location first [14].

Google Maps is a service provided entirely free by Google. By collecting data from sources such as satellites and ordinance survey sheets, they have created an accurate collage of worldwide maps, which is used globally by millions of businesses and individuals today. Google Maps also includes an Application Programming Interface (API), making map data and location based searching available to software developers for free. This has promoted a significant increase in the number of location based services available for mobile devices [14].

B. Increasing accuracy

Even with a location for the mobile device and an object ready to augment, there is still enough information to be able to render an object on screen in the correct position. There is no perspective in a flat image, so how does the mobile device know what parts of the building to render and to what scale it should be. GPS is only accurate to within 3 metres with a good signal. The position needs to be fine-tuned [14]

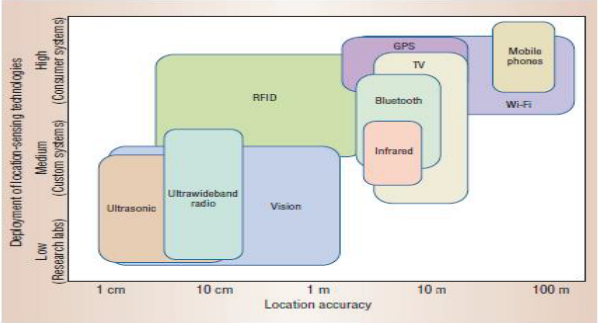


Fig. 9: Accuracy in methods of location retrieval [4]

XII. LOCATION SERVICES IN ANDROID OS:

Android gives your applications access to the location services supported by the device through classes in the android.location package. The central component of the location framework is the LocationManager system service, which provides APIs to determine the location of the underlying device. LocationManager is not instantiated directly. An instance of the class is created from the system by calling getSystemService(Context.LOCATION_SERVICE). The method returns a handle to a new LocationManager instance [7][8]. Figure 10 shows the representation of azimth pitch and roll as used by Android API methods.

A. Location Manager

This class provides access to the system location services. These services allow applications to obtain periodic updates of the device's geographical location, or to fire an application-specified Intent when the device enters the proximity of a given geographical location. The class cannot be instantiated directly but is retrieved through Context.getSystemService(Context.LOCATION_SERVICE) [9].

B. Location Class

Location: A data class representing a geographic location. A location can consist of latitude,longitude, timestamp, and otherinformation such as altitude and velocity. All locations generated by the LocationManager are guaranteed to have a valid latitude, longitude, and timestamp [10].

C. Geocoder Class

Geocoder: A class for handling geocoding and reverse geocoding. Geocoding is the process of transforming a street address or other description of a location into a latitude, longitude coordinate. Reverse geocoding is the process of transforming a latitude, longitudecoordinate into a partial address. The amount of detail in a reverse geocoded location description may vary, for example one might contain the full street address of the closest building, or one might contain only a city name and postal code [11].

D. Address Class

Address: A class representing an Address, it is a set of Strings describing a location. The android.Location package contains classes that define Android location-based and related services. Address is one of the classes from this package. This class provides various methods to retrieve the country, latitude, longitude, locality, postal code. Table 4.1 illustrates some of the most useful functions [12].

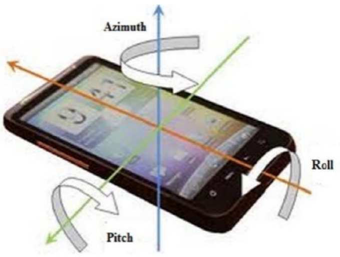


Fig 10 Representation of azimuth pitch and roll as used by Android API methods.

E. USAGE OF GOOG LE MAP A PI KEY IN XML FILE

Using the Google Maps Android API, maps can be embedded into an activity as a frag ent with a XML snippet. To use th e Google ma p API, the Google map API key has been obtained. The Google map API key is highlighted in the code fragment below.

```
<com .google.android.maps.MapView
android:id="@+id/mapView"
android:layout_width="fill_parent" android:layout_height
="fill_parent" android:clickable="true"android:apiKey="02AIy
M6bbvaEGk2r vm1GXrwXHwZKWDldmqVj 98w" />
```

XIII. CHALLENGES FOR LOCATION-BASED EXPERIENCES

Location-based experiences are in their infancy and the technologies on which they build are diverse and still maturing. Unsurprisingly, significant challenges need to be addressed before they reach their potential. In particular, it is important to be aware of the limitations of the technologies involved.

- Dealing with the uncertainty of location sensing
- Dealing with uncertainty of connection
- Interoperability
- Social and organization challenges [6].

XIV. CONCLUSION

Location-based services on smartphones have had great success in the consumer market, providing useful functions such as finding nearby points of interest. Next-generation LBSs promise to deliver even more interactive services to users and create a huge knowledgebase of location-tagged information. The major technological drivers of this are push notifications; better mobile network access through 3G and Wi-Fi; integration of advanced sensors on smartphones into applications such as accelerometers, digital compasses and still/video cameras; and Web 2.0 collaboration. As a result, analysts have predicted massive growth in the LBS market over the next few years. Today’s mobile application ecosystem allows users to download mobile applications ubiquitously.

The future of LBS in the both consumer and enterprise arenas promises to be very exciting; achieving the ultimate goal of true augmented reality based context-aware computing may not be far away.[2]

REFERENCE:

[1] “Dieter Schmalstieg, Mobile Computing Meets Augmented Reality”, Graz University of Technology, Austria http://igd-r.fraunhofer.de/fileadmin/jubilaeum2012/Vortraege-SmB-Forum/2012-11-14_Fraunhofer-IGD-25_Schmalstieg_web.pdf accessed on 25/5/2014

[2] Sidney Shek, “Next-Generation Location-Based Services for Mobile Devices”, CSC Grants February 2010

[3] Natalie Jun Pei Chin, Critical Success Factors of Location-Based Services, University of Nebraska – Lincolnhttp://digitalcommonsa.unl.edu/busineaccessed on 25/5/2014

[4] Mike Hazas, James Scott, and John Krumm. “Location-aware computing comes of age”, IEEE Computer Magazine, pages 95-97, February 2004

[4] Stefan Steiniger, Moritz Neun and Alistair Edwardes, “Foundations of Location Based Services”Accessed on 20/5/2014

[5]Reichenbacher, T., 2004,“Mobile Cartography - Adaptive Visualisation of Geographic Information on Mobile Devices”(PhD)

[6] Steve Benford, “Future Location-Based Experiences, School of Computer Science & IT,The University of Nottingham”, JISC Technology and Standards Watch

[7] Sudeshna Mukherjee, “Local Points of Interest Using Augmented Reality”, spring 2013, accessed on 20/5/2014

[8]GOOGLE, Location and maps. Android, http://developer.android.com/guide/topics/location/index.html#maps, accessed May2014, n.d.

[9] GOOGLE, LocationManager. Android, http://developer.android.com/reference/android/location/LocationManager.html, accessed December 2012, n.d.

[10]GOOGLE, Location. Android, http://developer.android.com/reference/android/location/Location.html, accessed December 2012, n.d.

[11]GOOGLE, GeoCoder. Android, http://developer.android.com/reference/android/location/Geocoder.html, accessed December 2012, last modified April 2013.

[12]GOOGLE, Address. Android, http://developer.android.com/reference/android/location/Address.html, accessed December 2012, last modified April 2013.

[13] Thomas Olsson,“User Expectations and Experiences of MobileAugmented Reality Services” Tampere University of Technology Tampere 2012, ISBN 978-952-15-2931-3

[14] Paul David Clegg, “A Location Based Service for Architecture, Using Augmented Reality”.

[15] www.flickr.com

[16] www.yelp.com

Er. Mrs. Sagaya Aurelia(November 9,1978) par-time research scholar in Bharathidasan university . Now she is with department of Computer Science, Faculty of Education, Azzaytuna University, Bani-walid, Libya. She received her Diploma in Electronics and Communication (1997),B.E (Bachelor of Engineering specialized in Electronics and Communication Engineering(2000) and M.Tech in Information Technology(2004),she has alsodoneherPostgraduation diplomas in Business Administration (PGDBA) and Journalism and Mass Communication(PGDJMC). She has received Brainbench certification in HTML. Her current research interest includes Virtual reality, augmented reality and Human Computer Interaction and User interface Design. She has authored14 papers and attendance several national and international level workshops and conferences.

Dr. Durai Raj is currently working as Assistant Professor, School of Computer Science, Engineering and Applications, Bharathidasan University, Tiruchirappalli, Tamilnadu, India. He completed his Ph.D. in Computer Science as a full time research scholar at Bharathidasan University on April, 2011. He received master degree (M.C.A.) in 1997 and bachelor degree (B.Sc. in Computer Science) in 1993 from Bharathidasan University. Prior to this assignment of Assistant Professor in Computer Science at Bharathidasan University, he was working as a Research Associate at National Research Centre on Rapeseed-Mustard (Indian Council of Agricultural Research), Rajasthan, and as a Technical Officer (Computer Science) at National Institute ofAnimal Nutrition and Physiology (ICAR), Bangalore for 12 years. He has published 26 research papers in both national and international journals. His areas of interest include Artificial Neural Network, Soft Computing, Rough Set Theory and Data Mining.

Dr. Omer Saleh Mahmod Jamah (January 25,1973) is now the Director of Post graduate cum Research and Development and Head of the department of Computer science, Faculty of education, Azzaytuna university, Baniwalid, Libya. He received his B.Sc. in Control System and Measurement (1995), M.Sc. in Electrical and Computer Measurement (2004), and Ph.D. in Electrical engineering, Automatics computer science and electronics from AGH University of technology, Krakow, Poland. He has done his Diploma in Planning and time management from Canada Global Centre, Canada. Now he is heading Computer Science department, Faculty of Education, Azzaytuna University, Baniwalid, Libya. His research interest includes multicriteria optimization for solving optimal control problems and Fuzzy logic. He has published 12 papers and attended various national and international Level conferences and workshops.

Graph Traversal on One-Chip MapReduce Architecture

Voichița Dragomir

***Abstract*—** A matrix based algorithm on a MapReduce engine is described for the graph traversal problem. The envisaged architecture is implemented as a one-chip many-core structure. The MapReduce architecture of the engine used to implement the algorithm is described. Both, dense and sparse matrices are considered. For evaluation a Scheme based simulator is used.

***Keywords*—**parallel computing, MapReduce, many-core, graph traversal, parallel algorithm.

I. Introduction

PROCESSING large graphs is becoming increasingly important for many domains. The ability to efficiently search large-scale graphs is becoming more and more important as we seek to model Internet-scale phenomena. Graphs are important in various fields like Computer Science, Biology, Chemistry and Social Sciences. Fundamental to graph search applications is graph traversal, a process of visiting all of the vertices and edges in the graph.

Parallel breadth-first search (BFS) algorithms have been implemented on multi-core processors, which are still based on the shared-memory model. They are limited in the number of cores, the memory size and they are non-scalable for big data size [1] [2].

Another implementation is done in cloud, where the MapReduce approach is limited by the latency introduced by the communication network [3] [4].

In this paper we present an approach for implementing BFS scheme, the common graph traversal algorithm, based on the MapReduce framework.

What is new in our approach is that we are going to achieve this on a one chip many-core structure not on multi-core or distributed computing, which implies multithreading, dividing the problem into threads and processing them simultaneous on multiple cores. These techniques have its limitations due to the communication and power issues.

So, we propose a new approach of this algorithm.

The system we work with is a *many-core* chip with a *MapReduce* architecture that performs best on matrix-vector operations.

Therefore, we designed an algorithm based on dense and sparse matrix operations.

The current existing solutions are shown in Fig. 1.

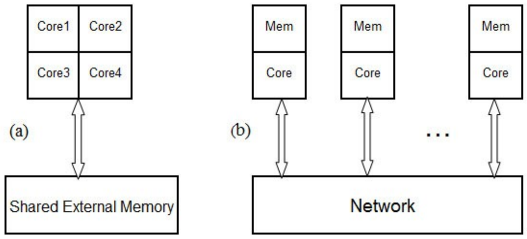


Fig.1. Current solutions: a) Multi-core architecture
b) Cloud MapReduce architecture

The problem is that both of them have limitations:
a) the cores compete for the shared resource (the external memory) known as the bottleneck effect
b) the individual cores have access to their own memory but there is a latency in communicating between the machines, so there is a significant increase in energy and time use.

Our solution is shown in Fig. 2:

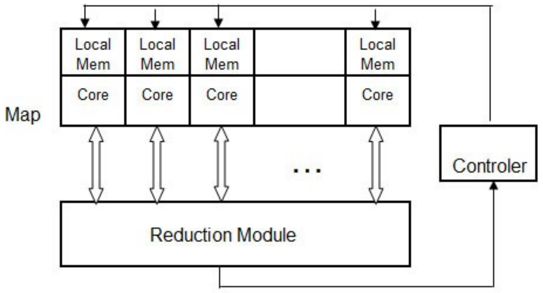


Fig.2. One-chip MapReduce architecture.

This structure has a very short response time because of the controller and the *log*-depth Reduce Module that works for many (thousands) cores. The main features of this structure are: high degree of parallelism, the cores are small and simple, the local memory is big enough for data mining applications.

There is another one-chip MapReduce approach based on the Intel SCC family. In [5] and [6] two different MapReduce applications are presented. The use of this general purpose array of processors has a much slower response because it has no more than 48 cores (which are much too complex for

This work has been funded by the Sectoral Operational Program Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/132397.

Voichita Dragomir is with The Politehnica University of Bucharest - Faculty of Electronics, Telecommunications and Information Technology, str. Iuliu Maniu 1-3. (e-mail: voichita.dragomir@upb.ro).

solving this kind of problem) and the MapReduce functionality is implemented in software, not hardware, as in our case.

In section II of this paper, we will present the one-chip MapReduce architecture used to solve the problem of breadth-first search (BFS) graph traversal. Section III presents the matrix based algorithms for both dense and sparse matrix. Section IV is the evaluation of these algorithms.

II. The MapReduce Architecture

The one-chip MapReduce architecture is defined by the data domains and the instruction set.

There are two data domains:

- V(vector) domain
- S(scalar) domain

V domain consists of the following vectors having the size equal with the number of cells:

indexVector $= < 0, 1, \dots n\text{-}1 >$ *// ix*
activeVector $= < a_0, a_1, \dots a_{n-1}>$
vector$_0$ $= < v_{00}, v_{01}, \dots v_{0n-1} >$
vector$_1$ $= < v_{10}, v_{11}, \dots v_{1n-1}>$
…
vector$_{p-1}$ $= < v_{p-10}, v_{11}, \dots v_{1n-1}>$

S domain is the external memory of the chip

scalarMemory $= < s_0, s_1, \dots s_{m-1}>$

The vector *indexVector* is used to identify the position of each cell in the linear array of cells. The vector *activeVector* is used to activate the cells. If the *i*-th component of the *activeVector* is 0, than the *i*-th cell is active, i.e., the current instruction is performed.

Instruction Set Architecture defines the operations performed over these two data domains. A short selection follows:

- `(ResetActive)`: activates all the components of the vectors involved in the operation that will be executed.
- `(Where x)`: keep active, from the active vector components, only the vector components where the Boolean vector x returns 1.
- `(ElseWhere)`: keep active only the vector components where the Boolean vector x returned 0 in the previously executed Where function.
- `(EndWhere)`: returns to the configuration of active components available before the execution of the previously executed Where function.
- `(SetVector x y)`: where the x is the vector's address in V, and the y expression returns the vector's content.
- `(SetStream x y)`: where x is the start address in S and y is the stream's content.
- `(BinaryOp x y)`: where BinaryOp = {Add, Sub, Mult,...}, and the expressions x, y is any combination scalar and/or vector operands
- `(UnaryOp x)`: where UnaryOp = {Inc, Dec, Clear, Abs, ...}, and x is the scalar or vector operand.
- `(RedOp x)`: where RedOp = {RedAdd, RedMax, RedOr, FirstIndex, ...} are the reduction operations and x is the vector's content, while RedOp returns a scalar.
- `(Search x y)`: search the scalar x in the active cells of the vector y; only the cells where the vector y has the value x remain active.
- `(KeepFirst x)`: where the x is the vector's address in V, while the function returns the vector with 0s in all positions except the first active position which remains unchanged.

A simulator for this MapReduce architecture written in DrRacket (a Scheme like programming language) is available at [7]. It was used to evaluate the algorithms developed in the next section.

III. The Algorithms

We chose an algorithm based on representing the graph as matrix, because the MapReduce architecture works very efficient on arrays. We are considering both the dense and the sparse matrix cases.

The algorithm we propose clears in each step a number of unnecessary edges from the graph in order to obtain the final tree, having the root in the starting vertex we chose.

We consider our graph has *n* vertices. Therefore the dense matrix used for representing the graph is *n* × *n* and it is stored in *n* vectors *v(0)... v(n-1).*

```
 // Graf traversal on MapReduce
   // vector(0): v(0)
   // ...
   // vector(n-1): v(n-1)
   // vector(n): t  // traceVec
   // active: a     // activeVector
   // index: ix     // indexVector
   // counter: c    // scalar
   // pointer: p    // pointer to current vector

      (SetStream c 1);
      (SetStream p startingVertex);
      (ResetActive)  ;
      (SetVector t (0 0 ... 0));
      (Where (ix = p));              // a = <11...101...1>
        (SetVector t   1)  ;
      (EndWhere)          ;          // a = <00...0>
     while (t contains 0)
       do while (t contains c)
           do(Search c t) ;
              (SetStream p (FirstIndex));
              (Where ((t = 0) & (v(p) = 1)))  ;
                (SetVector t (Inc c))   ;
                (EndWhere)   ;
              (SetVector v(p)0);// v(p) will be 0, where active
              (Where (ix = p)) ;
                (SetVector t x); // the first c in t is removed
                (EndWhere)   ;
          (Where (t=x))      ;
            (SetVector t  c) ;    // restore c
          (ElseWhere)        ;
          (SetStream c (Inc c));
      (ResetActive)          ;     // a = (0 0 ... 0)
for (i = 0; i < n; i = i+1) (KeepFirst v (i)) ;
```

The algorithm is explained step by step through the example shown in Fig. 3:

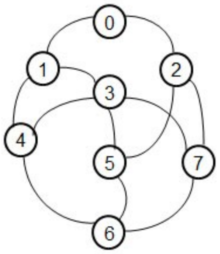


Fig.3. The graph we considered as example

A. Dense matrix algorithm

The initial state of the system is the following:

```
// Initial State
    0 1 2 3 4 5 6 7

v0  0 1 1 0 0 0 0 0
v1  1 0 0 1 1 0 0 0
v2  1 0 0 0 0 1 0 0
v3  0 1 0 0 1 1 0 1
v4  0 1 0 1 0 0 1 0
v5  0 0 1 1 0 0 1 0
v6  0 0 0 0 1 1 0 1
v7  0 0 0 1 0 0 1 0

c   = 1                // counter
t   = 0 0 0 1 0 0 0 0 // traceVector
a   = 0 0 0 0 0 0 0 0 // activeVector
fix = -                // firstIndex
```

where the vectors v0 to v7 represent the matrix, the counter c is initialized to 1, the trace vector t is initialized with 1 on the column of the initial vertex, and the activeVector a is full of 0s.

```
// The Main Steps
S1 Where((t=0)&(v3=1)) t<=c+1
   t= 0 2 0 1 2 2 0 2
   Clr(v3)
   c <= c+1  // c = 2
            *
      0 1 2 3 4 5 6 7

   v0 0 1 1 0 0 0 0 0
   v1 1 0 0 1 1 0 0 0
   v2 1 0 0 0 0 1 0 0
   v3 0 0 0 0 0 0 0 0
   v4 0 1 0 1 0 0 1 0
   v5 0 0 1 1 0 0 1 0
   v6 0 0 0 0 1 1 0 1
   v7 0 0 0 1 0 0 1 0
```

```
S2 Where((t=0)&(v1=1)) t<=c+1
   t= 3 2 0 1 2 2 0 2
   Clr(v1)
            *
      0 1 2 3 4 5 6 7

   v0 0 1 1 0 0 0 0 0
   v1 0 0 0 1 0 0 0 0
   v2 1 0 0 0 0 1 0 0
   v3 0 0 0 0 0 0 0 0
   v4 0 1 0 1 0 0 1 0
   v5 0 0 1 1 0 0 1 0
   v6 0 0 0 0 1 1 0 1
   v7 0 0 0 1 0 0 1 0
```

```
S3 Where((t=0)&(v4=1)) t<=c+1
   t= 3 2 0 1 2 2 3 2
   Clr(v4)
            *
      0 1 2 3 4 5 6 7

   v0 0 1 1 0 0 0 0 0
   v1 0 0 0 1 0 0 0 0
   v2 1 0 0 0 0 1 0 0
   v3 0 0 0 0 0 0 0 0
   v4 0 0 0 1 0 0 0 0
   v5 0 0 1 1 0 0 1 0
   v6 0 0 0 0 1 1 0 1
   v7 0 0 0 1 0 0 1 0
```

```
S4 Where((t=0)&(v5=1)) t<=c+1
   t= 3 2 3 1 2 2 3 2
   Clr(v5)
            *
      0 1 2 3 4 5 6 7

   v0 0 1 1 0 0 0 0 0
   v1 0 0 0 1 0 0 0 0
   v2 1 0 0 0 0 1 0 0
   v3 0 0 0 0 0 0 0 0
   v4 0 0 0 1 0 0 0 0
   v5 0 0 0 1 0 0 0 0
   v6 0 0 0 0 1 1 0 1
   v7 0 0 0 1 0 0 1 0
```

```
S5 Where((t=0)&(v7=1)) t<=c+1
   t= 3 2 3 1 2 2 3 2
   Clr(v7)
   c <= c+1   // c = 3
            *
      0 1 2 3 4 5 6 7

   v0 0 1 1 0 0 0 0 0
   v1 0 0 0 1 0 0 0 0
   v2 1 0 0 0 0 1 0 0
   v3 0 0 0 0 0 0 0 0
   v4 0 0 0 1 0 0 0 0
   v5 0 0 0 1 0 0 0 0
   v6 0 0 0 0 1 1 0 1
   v7 0 0 0 1 0 0 0 0
```

```
S6 Where(t=2) deactivate
   t= 3 2 3 1 2 2 3 2
        *   * * *   *
      0 1 2 3 4 5 6 7

   v0 0 1 1 0 0 0 0 0
   v1 0 0 0 1 0 0 0 0
   v2 1 0 0 0 0 1 0 0
   v3 0 0 0 0 0 0 0 0
   v4 0 0 0 1 0 0 0 0
   v5 0 0 0 1 0 0 0 0
   v6 0 0 0 0 1 1 0 1
   v7 0 0 0 1 0 0 0 0
```

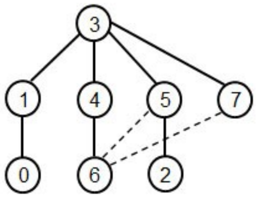
```
S7 Where((t=0)&(v0=1)) t<=c+1
   t= 3 2 3 1 2 2 3 2
   Clr(v0)
        *   * * *   *
      0 1 2 3 4 5 6 7

   v0 0 1 0 0 0 0 0 0
   v1 0 0 0 1 0 0 0 0
   v2 1 0 0 0 0 1 0 0
   v3 0 0 0 0 0 0 0 0
   v4 0 0 0 1 0 0 0 0
   v5 0 0 0 1 0 0 0 0
   v6 0 0 0 0 1 1 0 1
   v7 0 0 0 1 0 0 0 0
```

```
S8 Where((t=0)&(v2=1)) t<= c
   t= 3 2 3 1 2 2 3 2
   Clr(v2)
        *   * * *   *
      0 1 2 3 4 5 6 7

   v0 0 1 0 0 0 0 0 0
   v1 0 0 0 1 0 0 0 0
   v2 0 0 0 0 0 1 0 0
   v3 0 0 0 0 0 0 0 0
   v4 0 0 0 1 0 0 0 0
   v5 0 0 0 1 0 0 0 0
   v6 0 0 0 0 1 1 0 1
```

```
S9 Where((t=0)&(v6=1)) t<= c
   t= 3 2 3 1 2 2 3 2
   Clr(v6)
        *   * * *   *
      0 1 2 3 4 5 6 7

   v0 0 1 0 0 0 0 0 0
   v1 0 0 0 1 0 0 0 0
   v2 0 0 0 0 0 1 0 0
   v3 0 0 0 0 0 0 0 0
   v4 0 0 0 1 0 0 0 0
   v5 0 0 0 1 0 0 0 0
   v6 0 0 0 0 1 1 0 1
   v7 0 0 0 1 0 0 0 0
```

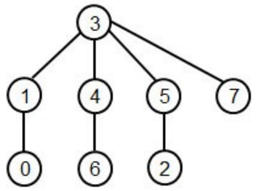
```
S10 Where(t=3) deactivate
    t= 3 2 3 1 2 2 3 2
       * * * * * * * *
       0 1 2 3 4 5 6 7

    v0 0 1 0 0 0 0 0 0
    v1 0 0 0 1 0 0 0 0
    v2 0 0 0 0 0 1 0 0
    v3 0 0 0 0 0 0 0 0
    v4 0 0 0 1 0 0 0 0
    v5 0 0 0 1 0 0 0 0
    v6 0 0 0 0 1 1 0 1
    v7 0 0 0 1 0 0 0 0
```

```
S11 KeepFirst in all vectors
      * * * * * * * *
      0 1 2 3 4 5 6 7

   v0 0 1 0 0 0 0 0 0
   v1 0 0 0 1 0 0 0 0
   v2 0 0 0 0 0 1 0 0
   v3 0 0 0 0 0 0 0 0
   v4 0 0 0 1 0 0 0 0
   v5 0 0 0 1 0 0 0 0
   v6 0 0 0 0 1 0 0 0
   v7 0 0 0 1 0 0 0 0
```



```
// The Final Result

      0 1 2 3 4 5 6 7

   v0 0 1 0 0 0 0 0 0
   v1 0 0 0 1 0 0 0 0
   v2 0 0 0 0 0 1 0 0
   v3 0 0 0 0 0 0 0 0
   v4 0 0 0 1 0 0 0 0
   v5 0 0 0 1 0 0 0 0
   v6 0 0 0 0 1 0 0 0
   v7 0 0 0 1 0 0 0 0
```

```
// The Tree
```



We keep only one edge on each line because if there are more 1s on a line it means there are multiple possible solutions, so we have to choose only one (it doesn't matter which one we keep; we choose the first one, because we have an instruction for this: (KeepFirst x)). So in this case, we can see, at step S10, the vertex6 has connections with vertex4, vertex5, and vertex7. So, we keep only the connection between vertices 4 and 6.

The final result consists of a matrix containing only seven 1s that means only seven edges. So, if we draw the resulting image there will be a tree, with 8 vertices and 7 edges.

B. Sparse matrix version

For the sparse matrix version we follow the algorithm from the previous subsection. So we start with the same matrix and start vertex.

	0	1	2	3	4	5	6	7
v0	0	1	1	0	0	0	0	0
v1	1	0	0	1	1	0	0	0
v2	1	0	0	0	0	1	0	0
v3	0	1	0	0	1	1	0	1
v4	0	1	0	1	0	0	1	0
v5	0	0	1	1	0	0	1	0
v6	0	0	0	0	1	1	0	1
v7	0	0	0	1	0	0	1	0

The sparse matrix representation is done by 3 vectors: one for edges, `e` (has the value 1 always), and the others for the location, one for the line, `l`, and another for the column of the respective edge, `c`. Revisiting the previous results, the vectors are:

```
e = <1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1> // edge
l = <0 0 1 1 1 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7> // line
c = <1 2 0 3 4 0 5 1 4 5 7 1 3 6 2 3 6 4 5 7 3 6> // column
```

```
step0 // pointer p = 3
a= <0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0>
e= <1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1>
l= <0 0 1 1 1 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7>
c= <1 2 0 3 4 0 5 1 4 5 7 1 3 6 2 3 6 4 5 7 3 6>

step1 // clear e where a=0 & l=3;
      //deactivate column 3
a = <0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0>
e = <1 1 1 1 1 1 1 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1>
l = <0 0 1 1 1 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7>
c = <1 2 0 3 4 0 5 1 4 5 7 1 3 6 2 3 6 4 5 7 3 6>

STAGE2, counter = 2. clear lines 1,4,5,7

step2  // clear e where a=0 & l=1
a = <0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0>
e = <1 1 0 1 0 1 1 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1>
l = <0 0 1 1 1 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7>
c = <1 2 0 3 4 0 5 1 4 5 7 1 3 6 2 3 6 4 5 7 3 6>

step3 // clear e where a=0 & l=4
a = <0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0>
e = <1 1 0 1 0 1 1 0 0 0 0 0 1 0 1 1 1 1 1 1 1 1>
l = <0 0 1 1 1 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7>
c = <1 2 0 3 4 0 5 1 4 5 7 1 3 6 2 3 6 4 5 7 3 6>

step4 // clear e where a=0 & l=5
 a = <0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0>
 e = <1 1 0 1 0 1 1 0 0 0 0 0 1 0 0 1 0 1 1 1 1 1>
 l = <0 0 1 1 1 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7>
 c = <1 2 0 3 4 0 5 1 4 5 7 1 3 6 2 3 6 4 5 7 3 6>

step5 // clear e where a=0 & l=7
 a = <0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0>
 e = <1 1 0 1 0 1 1 0 0 0 0 0 1 0 0 1 0 1 1 1 1 0>
 l = <0 0 1 1 1 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7>
 c = <1 2 0 3 4 0 5 1 4 5 7 1 3 6 2 3 6 4 5 7 3 6>

step6 // inactivate columns 1,4,5,7
 a = <1 0 0 1 1 0 1 1 1 1 1 1 1 0 0 1 0 1 1 1 1 0>
 e = <1 1 0 1 0 1 1 0 0 0 0 0 1 0 0 1 0 1 1 1 1 0>
 l = <0 0 1 1 1 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7>
 c = <1 2 0 3 4 0 5 1 4 5 7 1 3 6 2 3 6 4 5 7 3 6>

STAGE3, counter = 3. clear lines 0,2,6

Step7 // clear e where a=0 & l=0
 a = <1 0 0 1 1 0 1 1 1 1 1 1 1 0 0 1 0 1 1 1 1 0>
 e = <1 0 0 1 0 1 1 0 0 0 0 0 1 0 0 1 0 1 1 1 1 0>
 l = <0 0 1 1 1 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7>
 c = <1 2 0 3 4 0 5 1 4 5 7 1 3 6 2 3 6 4 5 7 3 6>

step8  // clear e where a=0 & l=2
 a = <1 0 0 1 1 0 1 1 1 1 1 1 1 0 0 1 0 1 1 1 1 0>
 e = <1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 0 1 1 1 1 0>
 l = <0 0 1 1 1 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7>
 c = <1 2 0 3 4 0 5 1 4 5 7 1 3 6 2 3 6 4 5 7 3 6>

step9 // clear e where a=0 l=6
 a = <1 0 0 1 1 0 1 1 1 1 1 1 1 0 0 1 0 1 1 1 1 0>
 e = <1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 0 1 1 1 1 0>
 l = <0 0 1 1 1 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7>
 c = <1 2 0 3 4 0 5 1 4 5 7 1 3 6 2 3 6 4 5 7 3 6>

STAGE4, we let only the first on each line (because
we have multiple solutions)

step10// let only the first l=6
 a = <1 0 0 1 1 0 1 1 1 1 1 1 1 0 0 1 0 1 1 1 1 0>
 e = <1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 0 1 0 0 1 0>
 l = <0 0 1 1 1 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7>
 c = <1 2 0 3 4 0 5 1 4 5 7 1 3 6 2 3 6 4 5 7 3 6>

THE FINAL REZULT (There are only 7 edges left, the
same result as in the first version)

e= <1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 0 1 0 0 1 0>
```

IV. Evaluation

The dense matrix version of the algorithm provides, for a *n*-vertex graph, the execution times *t(n)* in the following two limit cases:

- the resulting tree has the minimum number of levels:

$$t_{min}(n) = 24.5n - 2$$

because the main `while` loop (see the algorithm in the previous section) is executed only once

- the resulting tree has the maximum number of levels:

$$t_{max}(n) = 34.5n – 22$$

because the main `while` loop is executed *n* times.

In both cases

$$t(n) \in O(n)$$

while the time per vertex is in

$$O(1).$$

For $n = 1024$, the actual execution time results in the time interval of *(24.5–34.5)* clock cycles per vertex, depending on the depth of the resulting tree.

The previously described computation must be considered as being performed on a MapReduce parallel processor having the following performances, measured on a *65 nm* actual implementation [8]:

- *100 GOPS/Watt* (Giga Operations Per Second / Watt)
- *5 GOPS/mm^2*,

while the current sequential engines (x86 architecture) have, in the same technology:

- *~ 1 GOPS/Watt*
- *~ 0.25 GOPS/mm^2*

V. CONCLUSIONS

The current sequential algorithms provide execution time in $O(n+m)$, where n is the number of vertices, and m represents the number of edges. Our approach provides time in $O(n)$ using an architecture with *100*x performance per *Watt* and *20*x performance per mm^2 compared with the current technologies.

ACKNOWLEDGMENT

This work has been funded by the Sectoral Operational Program Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/132397.

REFERENCES

[1] G. Revesz, “Parallel Graph-Reduction With A Shared Memory Multiprocessor System,” *IEEE Computer Languages,* 1990, New Orleans, LA, pp. 33-38.

[2] M. Yasugi, T. Hiraishi, S. Umatani, T. Yuasa, "Dynamic Graph Traversals for Concurrent Rewriting using Work-Stealing Frameworks for Multi-core Platforms", *IEEE Conference on Parallel and Distributed Systems* (ICPADS), 16th edition*,* 2010, pp. 406 - 414.

[3] M. Cosulschi, A. Cuzzocrea, R. De Virgilio, "Implementing BFS-based Traversals of RDF Graphs over MapReduce Efficiently." *IEEE Conference on Cluster, Cloud and Grid Computing* (CCGrid)*,* may 2013, Delft, pp. 569 - 574.

[4] Qian Lianghong, Fan Lei, Li Jianhua, “Implementing Quasi-Parallel Breadth-First Search in MapReduce for Large-Scale Social Network Mining” *IEEE Conference on Computational Aspects of Social Networks (CASoN)*, Fifth International Conference, 2013, pp. 7 - 14.

[5] A. Papagiannis, D.S. Nikolopoulos, “MapReduce for the Single-Chip-Cloud Architecture” *Institute of Computer Science (ICS), Foundation for Research and Technology – Hellas (FORTH)*,Greece, 2011.

[6] A. Tripathy, A. Patra, S. Mohan, R. Mahapatra, “Distributed Collaborative Filtering on a Single Chip Cloud Computer” *IEEE Conference on Cloud Engineering (IC2E),* 2013, pp. 140 - 145.

[7] connexAppl2013WEB.rkt. at http://www.anselm.edu/internet/compsci/faculty_staff/mmalita/HOMEPAGE/research.html

[8] Gheorghe Stefan: "One-Chip TeraArchitecture", in *Proceedings of the 8th Applications and Principles of Information Science Conference*, Okinawa, Japan, January 2009.

Voichița Dragomir teaches Digital Integrated Circuits at Politehnica University of Bucharest - Faculty of Electronics, Telecommunications and Information Technology. Her research interest is in parallel architecture and algorithms.

Cluster Head Influence based cooperative Caching in Wireless Sensor Networks

Ashok Kumar

Abstract— **Cooperative caching harnesses the combined data storage capacity of memory constrained Sensor Nodes (SNs) for data caching. There are a number of Wireless Sensor Network (WSN) applications where, sink node may require recent history data communicated by a particular SN. In such situation, the data request at the sink needs to be served in short latency and with minimal energy consumption. A cooperative caching protocol termed as Cluster Head Influence based cooperative Caching (CHIC) has been proposed as an effective and efficient technique to achieve these goals concurrently. The proposed cooperative caching scheme is based on the influence of cluster heads in clustered WSN. We have proposed new cache admission control, cache discovery and cache replacement protocols under the CHIC cooperative cache management.**

Index Terms—**Admission control, byte hit ratio, cache discovery, cache replacement, cluster head, query latency.**

I. Introduction

WSN field has emerged as an effective technology for distributed sensing and monitoring of a remote field/terrain which may not be accessible by conventional means of sensing/monitoring. One of the major drawbacks of WSNs is that SNs are highly energy constrained devices, having limited battery life and it is hard or rather impossible to rejuvenate/replace the batteries in inaccessible and inhospitable environments. Despite such constrains, WSNs are expected to perform their intended task of sensing/monitoring the sensing field for maximum possible time period. The short life span of WSN may lead to enhanced cost for continuous sensing/monitoring of the area of interest, as a lot of efforts and expenses are involved in frequent redeployment of SNs.

Majority of WSN applications are data centric which require transmission of sensed data to sink node situated inside/outside the sensing field, through single/multi hop transmission. Once the data is collected at sink, it preprocesses the data as per requirement of the application and further transmits to the end user via Internet or some other wireless/wired media. Due to presence of large number of SNs in the sensing field, vast amount of data is communicated to the sink during each sensing round resulting in large number of transmit/receive operations. As data transmission and reception operations are major sources of energy consumption in the WSN, considerable amount of energy could be preserved if number of data transmission/reception operations could be reduced. This could be achieved if data communication to the sink is performed only as per the request of the user/application via sink node. As the request pattern of data is hard to be predicted, sink node can request data from any region at any time. Therefore, SNs need to perform continuous periodic sensing to fulfill this objective and sensed data must be stored either by the same SN who has sensed it or by some other SN in the network, so that data could be provided to the sink node as and when request for such data is received. Since, the data storage capacity of the SN is limited, it can only store limited amount of data temporarily. However, collective data storage capacity of SNs combined together can be harnessed for data caching operation and sensed data is spread across distributed nodes throughout the sensor field.

There are numerous applications, such as tracking the movement of certain object/animal/vehicle, tracking the spread of certain phenomenon such as forest fire, creating panoramic view of an area by collecting the visual information transmitted by SNs equipped with micro-camera capable of capturing only a narrow view of area surrounding the node, etc., sink node may require recent history data reading along with present data readings transmitted by the SNs. This requirement can be easily fulfilled if sink node keeps storage of all history data it receives. But this requires very large storage capacity available at the sink, as hundreds of SNs keep on pumping sensed data continuously to the sink. Although, in many WSN protocols reported in literature, sink node has been assumed to have unlimited power, processing, and data storage resources available at its disposal; however, there may be a number of scenarios where such assumptions may not be valid and sink node may have limited resources.

Data caching is a well-researched field in wireless ad hoc networks and lot of caching protocols including optimized cache discovery, cache admission, cache replacement, and cache consistency processes are available in literature for such networks [1]-[11]. However, data caching is altogether a different and challenging task in WSNs due to (i) limited storage capacity of individual SNs, (ii) data request pattern (many-to-one data dissemination), and (iii) limited battery capacity of SNs. The fundamental objective of cooperative caching in WSN is to identify the SNs which will cache the data, deciding whether SN should cache the particular data or not, deciding which of the cached data of SN be deleted to accommodate new data, how long the data should be cached in the network, and deciding optimal path to get the query resolved with shortest possible latency and minimum communication overhead for energy efficient operation of WSN.

Considering the special requirements of WSNs to resolve above problems, we have proposed an energy efficient cooperative caching technique called Cluster Head Influence based

Ashok Kumar is working as Associate Professor E&CED, National Institute of Technology Hamirpur (HP) India-177 005. (Email: ashok@nith.ac.in)

cooperative Caching (CHIC), which effectively utilizes the clustering in WSN to optimize the query resolution. The proposed scheme utilizes the combined storage capacity of SNs present in the sensing field, thus increasing the size of cumulative caching. It is aimed to achieve low energy consumption per query response, low access latency, and creating optimal number of data copies in the network to avoid the wastage of cache space on one hand and provide enough replication of data for efficient query resolution on the other.

II. Network Model and Assumptions

Proposed scheme is designed for hierarchical WSNs in which SNs are divided into clusters. If it is to be used for flat network, the network first needs to be divided into clusters using any energy efficient clustering protocol available in literature. Our proposed scheme is directly applicable for such clustered networks where clusters and CHs required for cooperative cache implementation are already available in the network. Following assumptions are made to implement proposed caching scheme in WSN:

- N_T number of SNs are assumed to be deployed randomly in a sensing field of radius R with uniform node density ρ.
- SNs are static and are aware of their location coordinates.
- SNs are homogeneous and have same caching capacity.
- WSN is assumed to have been divided into clusters, each cluster having elected its own CH.
- All SNs communicate their sensed data to a single sink node situated at the center of circular sensing field.
- Data request (query) is initiated by the sink node only and requested data is always destined to the sink node i.e. all data request (query) paths originate at the sink and all data path (query resolution path) terminate at the sink only.
- Computational capability, data storage capacity, initial battery energy, and communication range of sink are higher than all other SNs deployed in the sensing field.
- The set of data items is denoted by $D = \{d_1, d_2, \ldots d_N\}$, N is the total number of data items and d_j $(1 \leq j \leq N)$ is a data identifier. D_i denotes the actual data for item with id d_i. Size of data item d_i is S_i. Data item can be originated from any SN.

III. Proposed Cooperative Caching Scheme

Cooperative caching protocols proposed in literature for WSN are based on the estimation of SN importance in given network topology and nodes are selected as data caching nodes, based on their relative importance. CHs in a hierarchical WSN are always selected based on their communication capability within their own cluster members [12]. As, all SNs communicate their sensed data through their respective CHs and further this data is relayed towards the sink via CHs only; this signifies the importance of CHs in data caching. In proposed protocol, a cluster head is allowed to cache data utilizing the combined storage capacity of its cluster members and has been termed as caching cluster head (CCH).

A. Cache Discovery Process

Cache discovery is the process by which sink locates the requisite data by broadcasting query related to the data item, within the sensor network. To avoid unnecessary communication query overhead, efficient and optimized cache discovery process is an essential requirement of any WSN cooperative caching protocol. Query initiated by sink is addressed in terms of data item id $D = \{d_1, d_2, \ldots d_N\}$, which essentially comprises of the sensing node id $N_{id} = \{d_1, d_2, \ldots d_T\}$ and time stamp $t_s = \{t_{s1}, t_{s2}, \ldots\}$. Time stamp t_s is the time at which the data is sensed. In the proposed protocol, it is assumed that whenever a SN senses an event, it communicates the sensed event to its CH by transmitting a data frame DF. The data frame comprises of SN id N_{id}, time stamp t_s, and TTL of the data item. Two SNs situated in different clusters can be assigned same node id N_{id} and since the process of assigning node id is cluster centric it is relatively easy due to small number of SNs in a cluster.

CHs may or may not carry out data aggregation/fusion operation on data received from their respective cluster members during one data sensing round. In any case, CH generates a data frame by adding its own cluster head id CH_{id} to the received frame from SNs, and relays it towards the sink. While this data frame traverse towards sink through various relay CHs on the way, each CH maintains a table of CH_{id} whose data they have routed. Since, each CH need to relay data of small number of CHs, the size of data table is quite small. However, this tabulation saves a lot of communication overheads during the query resolution process.

The proposed cache discovery process is illustrated in Fig. 1. Before sending query for a data item, sink first of all checks its own local cache. In case data item is found there and is valid, query is immediately served without broadcasting it any further. If the data item is not found in its cache, it broadcasts the query within its first hop CLUSTER called FH-CLUSTER with radio power R. All CHs residing within first hop CLUSTER search their local cache (local cache includes the cache of CH and its associated cluster members), if data is found within their local cache, query is served by sending the queried data item to the sink. If data item is not found in their cache, they search their routing table and check if such data was routed through them (simply by matching the CH_{id} of queried data item with CH_{id} in their respective routing table). Those CCHs for which CH_{id} of queried item does not match with any of CH_{id} in routing table, stop further broadcast of query. This process avoids query broadcast through those routes where data in not residing, thus saving considerable amount of energy in the network. Those CCHs which find routing information of queried data item in their respective routing table, broadcast the query to CCHs residing in CLUSTER with higher index (away from the sink). Those CCHs which are residing either in same CLUSTER or CLUSTER with lower CLUSTER index value (towards the sink) ignore the query. This process continues till the query is resolved. If queried data item is not available anywhere on the routing path, the query is served by the data originating node through its CH (as data item is always cached at originating node till expiry of data item TTL).

B. Cache Admission Control

When a node receives the data, the process of cache admission control decides whether the received data is to be cached or not. In WSN, each sensor node can store the received data in its local storage. In caching terminology, the sensor node, which stores the received data, is termed as caching node (CN). Caching of every received data at the CN is not possible as it requires large amount of memory at the CN and if proper and efficient caching decision is not made, it may lower the efficiency of caching scheme, leading to increase in data latency with lower probability of cache hits [8]-[11].

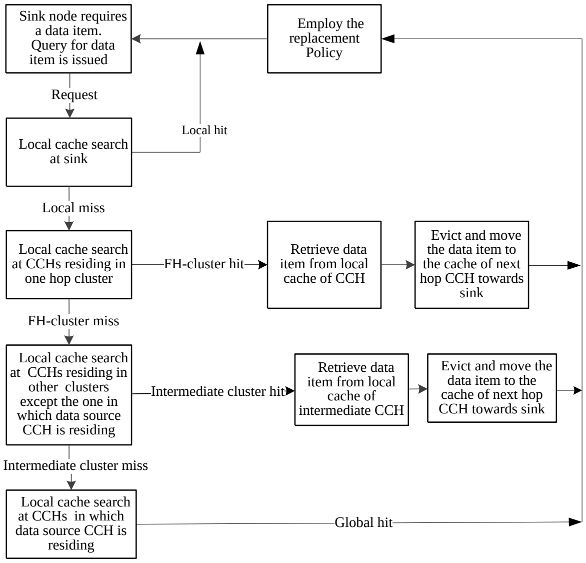

Sink node requires a data item. Query for data item is issued
Employ the replacement Policy
Request
Local hit
Local cache search at sink
Local miss
Local cache search at CCHs residing in one hop cluster
FH-cluster hit
Retrieve data item from local cache of CCH
Evict and move the data item to the cache of next hop CCH towards sink
FH-cluster miss
Local cache search at CCHs residing in other clusters except the one in which data source CCH is residing
Intermediate cluster hit
Retrieve data item from local cache of intermediate CCH
Evict and move the data item to the cache of next hop CCH towards sink
Intermediate cluster miss
Local cache search at CCHs in which data source CCH is residing
Global hit

Fig. 6.2: Cache discovery and data retrieval process in CHIC protocol.

In our proposed scheme, the caching decision is based on two prime questions (i) whether the CN is appropriate caching site to store a particular data, and (ii) whether a particular data is appropriate data for caching at a particular CN. Data is admitted to the cache only if appropriate data is received at appropriate CN. Fitness/appropriateness of a CN or a data item for cache admission control is decided by significance of data item and CN. In our protocol, we have used the appropriate CHs for caching known as Caching Cluster Head (CCH). Although, every CH can act as CCH due to higher significance compared to other SNs deployed in the network; however there may be difference in the significance of CHs due to variation in their cluster members, adjacent CHs, and distance from the sink. In proposed protocol significance of CH has been defined in terms of Cluster Head Significance Index (CHSI) as defined below:

- *Cluster Head Significance Index (CHSI)*

CHSI describes the probability of cluster head to act as Caching Cluster Head (CCH) in the network. It is measured based on following parameters:

i. Cluster Head Influence Index (CHII)

The influence of a CH in the network is measured in terms of its one hop degree of connectivity to its neighboring CHs. If a CH is surrounded by more number of CHs, probability of a query to traverse through such CHs is also high; therefore, efficient query resolution can be achieved if such CHs are selected as CCHs.

ii. Node Influence Index (NII)

NII of a CH is measured in terms of number of nodes present in its own cluster. If a CH is having more number of SNs in its cluster, it has the availability of more cache for data storage; as it can use collaborative cache of all its cluster members, along with its own cache. Availability of larger cache size within such cluster enhances the probability of query resolution by such CHs.

iii. Location Influence Index (LII)

LII of a cluster head is measured as its distance in terms of hops from the sink. In case of our scenario, WSN is having single sink, and the data request/query is always originated at the sink and if the sink is not able to resolve it from its own cache, it broadcasts it within the network. To achieve minimum latency and minimum energy expenditure for query resolution, the query must get resolved at nearest possible CH from the sink.

As $CHSI \propto (CHII) \cap (NII)$ and $CHSI \propto \frac{1}{LII}$. Therefore,

$$CHSI = \left\lfloor \frac{CHII \times NII}{LII} \right\rfloor \qquad (1)$$

CHSI provides a means for indexing the CH as per its fitness to act as CCH.

Other parameter which influences the cache admission protocol in WSN is fitness of a particular data item for caching at a particular CCH. Therefore, another parameter i.e. Data Popularity Index (DPI_i) has been defined to measure the fitness of data item d_i for its caching at a particular CCH.

- *Data Popularity Index (DPI)*

DPI is a measure and indexes the popularity of data at the sink. Such indexing of data popularity is useful for deciding the cache admission and cache replacement policy for an efficient cooperative caching scheme. In proposed protocol DPI has been based on following parameters:

i. Access Frequency of Data Item

Access frequency $f_i(t)$ of a data item d_i is a direct measure for data popularity of d_i at the sink. Frequently accessed data items are supposed to be more popular than other data items. Access frequency $f_i(t)$ for data item di in local cache of a CCH is given as:

$$f_i(t) = \frac{a_i}{\sum_{k=1}^{n} a_k} \qquad (2)$$

Where a_i is mean access rate of data item d_i over a time period T and n is total number of data requests generated at the sink over time T.

ii. Time to Live (TTL) of Data Item

TTL of a data item is a measure of time for which a particular data item di is valid. Data item is declared invalid after expiry of its TTL period. A data item is assumed to have high DPI value if TTL of that data item is high. TTL value for a particular data item is decided by the SN and is based on the application. The data sensing node adds a data sensing time stamp t_s and TTL value t_{TTL} while transmitting the data to its CH. Data validity in the network is defined by (t_s - t_{TTL}) value.

iii. Distance of CCH from the Data Source

Data item sensed by a particular SN is not cached by the CCH within its own cluster, as in the proposed protocol the data is always cached by the sensing node for a time window based on the TTL value of data item. Therefore, the data item d_i is assigned higher DPI as it moves away from the data originating node i.e. closer towards the sink. The popularity of a data item d_i at a CCH is proportional to $d_{(CCH)i}$, the distance of CCH in terms of hops from data originating source of data item d_i. Taking into consideration all above mentioned parameters, DPI of a particular data item di can be given as:

$$DPI_i = f_i(t) \times TTL_i \times d_{(CCH)i} \tag{3}$$

Based on the fitness/appropriateness function of CH and data item expressed by (1) and (3), a new cache admission control policy has been proposed which allows caching of suitable data at a suitable CCH so that data request/query get resolved with minimum possible communication overhead and least possible data latency.

The proposed cache admission control policy allows the SN to cache all data items it has sensed and the data item remains in the cache of sensing node till the TTL expiry of data item. The node which senses the data item has been named as Data Originator Node (DON). This policy is proposed so as to avoid failure of any data request resolution as in worst case scenario the data item should be available at least at DON. Other SNs and CH of same cluster under which DON is residing are not allowed to cache the data item originated within same cluster. To increase proximity of the data items nearer to sink, it is always better to start caching of data items nearer to sink [9]. However, WSN utilizes reverse multicast data transmission policy for sensed data, where data sensed by the SNs is forwarded to the sink; whereas, query follows the multicasts data transmission policy, where single sink node multicast the query within the network. This type of conflicting data transmission policy in WSN makes the accomplishment of efficient data admission control a difficult task compared to MANET. Following policy has been employed for proposed cache admission control:

1. Initially all CHs present in the sensor network calculate their own CHSI using (1) and communicate CHSI value to the sink. On completion of this step sink node receives $[CHSI_1, CHSI_2, CHSI_3, ... CHSI_k]$, where k is number of CHs in the sensing field.
2. Sink node calculates $[CHSI_1, CHSI_2, CHSI_3, ... CHSI_k]_{\max}$, the CHSI index having the maximum value and communicates the $CHSI_{max}$ value to the CHs. Apart from communicating $CHSI_{max}$ value, sink node also decides the threshold value of cluster head significance index ($CHSI_{threshold}$) and communicates it to the CHs. This value is used by CHs while deciding the cache admission of a particular data item.
3. On receiving $CHSI_{max}$ value, each CH calculates normalized cluster head significant index value. For CH_j, normalized CHSI value is calculated as:

$$\left(CHSI_j\right)_{norm} = \frac{CHSI_j}{CHSI_{\max}} \tag{4}$$

4. Cache admission decision for data item d_i, received by CH_j is based on the following logic:

$$\left(CHSI_j\right)_{norm} \geq CHSI_{threshold} \tag{5}$$

If the logic of (5) is true, only then CH can act as CCH for a particular data item d_i, otherwise it simply route the data item to its next hop CH. However, even if the (5) is true, it does not guarantee the caching of data item di by CH_j, as the CH needs to check the CHSI index of its next hop CH on routing path i.e. $CHSI_{j+1}$ as well as the DPI indices DPI_i^j and DPI_i^j. DPI_i^j signifies the importance/significance of data item d_i at CH_j. Therefore, the caching decision for CH_j is based on the following logic:

$$\left\lfloor DPI_i^j \cap \left(CHSI_j\right)_{norm} \right\rfloor \geq \left\lfloor DPI_i^{j+1} \cap \left(CHSI_{j+1}\right)_{norm} \right\rfloor \tag{6}$$

Equation (6) requires the value of DPI_i^j and DPI_i^{j+1} i.e. data popularity index for data item d_i both at CCH_j as well CCH_{j+1}. However, soon after the deployment of WSN, when SNs initially start communicating their sensed data to the sink node, there may not be any data request/query from the sink. This preliminary sensed data also needs to be cached by some CCH in the network. In absence of any information regarding the data access rate during preliminary stage of sensor network, the value of access frequency $f(t)_i$ for data item d_i is assumed to be unity. Subsequently when the sink node starts sending request/query for certain data item, the CCH receiving the request starts calculating the access frequency $f(t)_i$ for that data item It is intuitive that if (6) is satisfied, CCH_j is the most appropriated node to cache the data item d_i, as it has better caching fitness as compared to its next hop CCH i.e. CCH_{j+1}. In the event of (6) is not true, the CCH_j simply needs to route the data item to CCH_{j+1}, without caching.

However, if cache of CCH_{j+1} is full then the problem of data caching becomes trivial, as CCH_{j+1} has to evict less popular data from its cache.

C. Cache Replacement Policy

Cache replacement policy is required when a CCH attempts to cache the data item but its cache is full. In such event, CCH has to evict some existing data from its cache to accommodate the new data item. To accommodate new data item, there is no option other than evicting an existing data item from the cache. The victim must be selected based on policy which results in minimum loss of information and minimum overheads [9]. In proposed protocol, cache replacement is carried out based on Data Popularity Index (DPI) value calculated as per (3). The proposed cache replacement works as follows:

- Data item originated at a particular node is always cached at that node and remains in the cache of that node till the expiry of its TTL value.
- CCH immediately evicts a data item from its cache as soon as its TTL value expires.
- A cost based cache replacement policy is proposed for eviction of cached items from local cache of CCH. The cost is based on DPI value of the data item. CCH always keeps the track of maximum and minimum DPI value of data items present in its local cache. (i.e. DPI_{max} and DPI_{min}). DPI_{max} and DPI_{min} value is updated regularly on arrival/eviction of new data item. When a new data item arrives at the CCH, it calculates its DPI value and compares the calculated DPI value with DPI_{max} value of data items stored in its local cache. If the DPI value of arrived item is more than the DPI_{max} value, it initiates the process of cache replacement; otherwise, the data item is routed

to next CCH. In the event of cache replacement, CCH first replaces the item with minimum DPI value i.e. DPI_{min}. The evicted item is routed to neighboring CCH for storage if its TTL value is still valid as it may be requested by the sink.

TABLE I
SIMULATION PARAMETERS

Parameter	Default value	Range
Network diameter	100 meters	50~400 meters
Number of nodes	400	100~500
Initial Energy of node	2 Joule	
Data packet size (k)	100 byte	
Threshold distance (d0)	87 meters	
Cache size	800 KB	200~1400 KB
Time to live (TTL) of data	300 sec	
Data rate	10 Kbps	
Transmission range	10 meters	
Zipfian skewness parameter (z)	0.8	
Mean query generate time	5 sec	2~100 sec
E_elect	50 nJ/bit	
ϵ_{fs}	10 pJ/bit/m2	
ϵ_{mp}	0.00134 pJ/bit/m4	
E_{aggr}	5 nJ/bit/signal	

IV. PERFORMANCE EVALUATION

A. Simulation Parameters and Performance Metrics

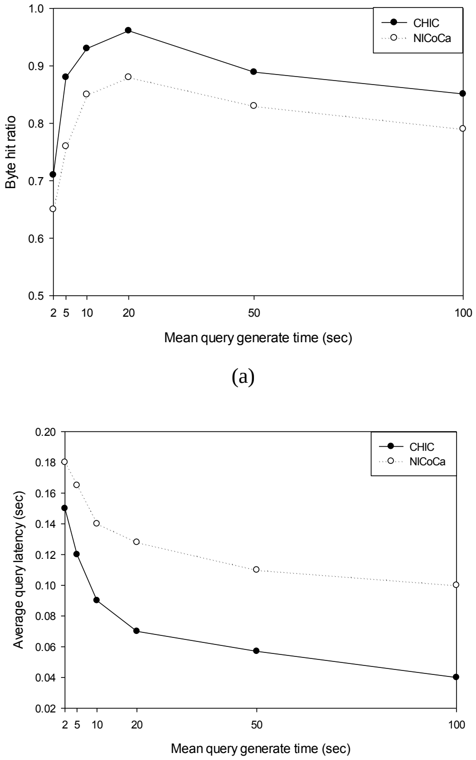
The simulation parameters are given in Table I. Performance of proposed protocol is compared with NICoCa [8] and is evaluated using following performance metrics:

- *Average query latency (T_{mean})*: The query latency is the time elapsed between transmission of query and its resolution at the sink. Average query latency (T_{mean}) is the query latency averaged over all the queries generated by the sink.
- *Byte hit ratio (B)*: It is the ratio of number of data bytes retrieved from the cache to the total number of requested data bytes. It is used as a measure of the efficiency of cache management.

B. Simulation Results

- *Effect of mean query generate time (T_q)*

The effect of mean query generate time (T_q) on byte hit ratio (B) and average query latency (T_{mean}) has been illustrated in Fig. 2. Mean query generate time is the time between two consecutive queries, averaged over total number of queries. T_q has been varied from 2 to 100 seconds. It can be seen from Fig. 2(a) that byte hit ratio increases with increase in T_q, reaches its maximum value around 20 and then again starts reducing. Initially, when the T_q value is very small, more number of queries are generated per unit time and very little time is available for settlement of cache to make required data items available at cache of CCHs near the sink. Therefore, all queries may not get resolved at local/FH-CLUSTER/intermediate CLUSTER cache and global hits are inevitable. As the T_q is increased, byte hit ratio starts increasing, since network has to resolve less queried per unit time, which gives enough time for settlement of network cache for providing highly queried data item at the cache of CCHs closer to the sink. This accounts for higher byte hit ratio. However, if Tq is further increased, it does not result in better hit ratio, in actual the hit ratio starts reducing. The reason behind this behavior is that, though enough time is available for cache to settle down between the queries, but due to expiry of TTL value of various cached data items, eviction process of these data items starts in the network, again destabilizing the balance of cache in the network, which results in lower byte hit ratio. Similar trends are observed for CHIC and NICoCa. Proposed CHIC protocol provides higher byte hit ratio compared to NICoCa due to better cache management policy proposed in CHIC. The cache admission control of CHIC ensures the availability of highly required data at CCH near the sink. Only small number of global hits are observed in case of CHIC protocol thus, proving the superiority.



(a)

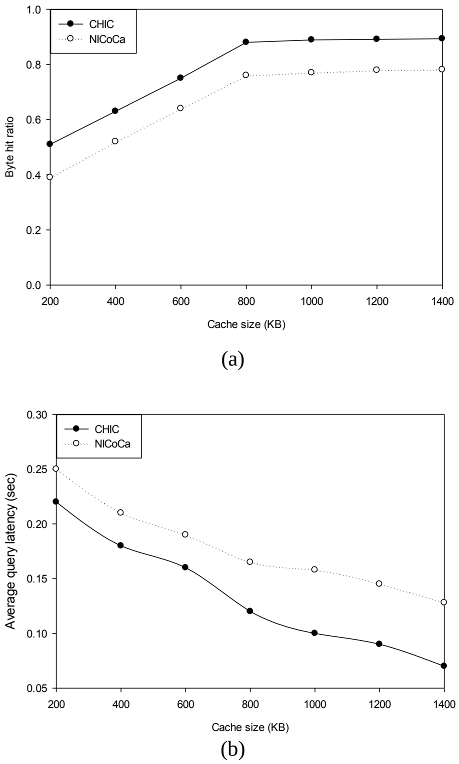
(b)

Fig. 2: Effect of mean query generate time on (a) byte hit ratio, and (b) average query latency.

Fig. 2(b) illustrates the variation in average query latency (T_{mean}) with T_q. It can be observed that at lower T_{mean}, the T_q value is high. It is due to the fact that generation of more number of queries per unit time results in more congestion and data collisions in the network thus, increasing the time in query resolution. Intuitively, T_{mean} reduces with increase in T_q. However, rate of decrease of T_{mean} keeps on reducing. This is due to the fact that if mean query generation time is very large, expiry of TTL value of large number of data items starts in between, resulting in small number of cache hit and in turns reduction in rate of decrease of T_{mean}. The cache management policy of proposed CHIC protocol ensures the availability of highly queried data items near the sink. Therefore, average query latency in proposed protocols is always less than NICoCa protocol, irrespective of the mean query generate time. Thus, CHIC outperforms the NICoCa in terms of average query latency.

- *Effect of cache size*

Fig. 3 illustrates the effect of cache size on byte hit ratio (B) and average query latency (T_{mean}). The cache size of SNs is varied from 200 KB to 1400 KB. It can be seen in Fig. 3(a) that initially the byte hit ratio increases with increases in the cache size. It is an obvious result, since higher size of cache helps the placement of data items at local/FH-CLUSTER/intermediate CLUSTER cache thus, increasing the probability of cache hits. However, the byte hit ratio becomes almost constant after about 800 KB cache size. This is due to the fact that in network, the data items start getting stale and evicted from the cache, due to expiry of their TTL. As the rate of data sensing in the sensing field is assumed constant, therefore an appropriate value of cache is enough to maintain the optimum byte hit ratio. Increasing the cache size beyond this limit does not add substantially to the byte hit ratio. It can be seen that proposed CHIC protocol provides better byte hit ratio compared to the NICoCa protocol.



(a)

(b)

Fig. 3: Effect of cache size on (a) byte hit ratio, and (b) average query latency.

Fig. 3(b) illustrates the effect of cache size on T_{mean}. Initially, when the cache size is small, T_{mean} is high, because due to smaller cache availability the queried data items may be available away from the sink. The data item availability nearer to the sink is better in CHIC protocol as can be seen by the lower value of T_{mean} compared to NiCoCa. With increase in cache size, T_{mean} starts reducing. Initially, the rate of decrease is high which reduces with further increase in the cache size. This behavior is attributed to the expiry of TTL value of data items and their eviction from the cache. Therefore, to maintain certain minimum standard value of T_{mean} and B, minimum value of cache size (800 KB in our case) is required. Increasing the cache size beyond this is futile. Byte hit ratio and average query latency of proposed CHIC protocol is better than NICoCa protocol irrespective of SN cache size. It is attributed to the efficient cache management policy adopted in the proposed protocol, which ensures the availability of highly queried data items closer to the sink; thus, enhancing the byte hit ratio and reducing the average query latency.

V. Conclusion

Proposed cooperative caching scheme CHIC exploits the influence of CHs already present in the sensing field. In CHIC protocol, CHs are used as CCH to cache the data. They can efficiently use the caching space of all SNs present in their respective clusters. In clustered network all data items are routed through the CH, this helps in efficient query resolution. We have proposed new cache admission control, cache discovery and cache replacement protocols under the CHIC cooperative cache management. The simulation results demonstrate the effectiveness of proposed protocol compared to similar caching protocol NICoCa.

REFERENCES

[1] G B. Raj, S.K. Vishvakarma, A.K. Saxena and S. Dasgupta, “Techniques for Low Power SRAM Design,” National Conference on Design Techniques for Modern Electronic Devices, VLSI & Communication Systems, pp. 55-61, 2007.

[2] A. Dunkels, J. Alonso and T. Voigt, “Distributed TCP Caching for Wireless Sensor Networks,” Annual Mediterranean Ad-Hoc Networks Workshop, pp. 1-17, 2004.

[3] A. Ayadi, P. Maille and D. Ros, “Improving Distributed TCP Caching for Wireless Sensor Networks,” IFIP Annual Mediterranean Workshop, pp. 1-6, 2010.

[4] K. Prabh and T. Abdelzaher, “Energy-Conserving Data Cache Placement in Sensor Networks,” ACM Transactions on Sensor Networks, Vol. 1, No. 2, pp. 178–203, 2005.

[5] S. Gupta, S. Mittal, S. Dasgupta and A. Mittal, “MIMO Systems For Ensuring Multimedia QoS Over Scarce Resource Wireless Networks," ACM International Conference On Advance Computing, 2008.

[6] J. Xu, K. Li, Y. Shen and J. Liu, “An Energy-Efficient Waiting Caching Algorithm in Wireless Sensor Network,” International Conference on Embedded and Ubiquitous Computing, Vol. 1, pp. 323-329, 2008.

[7] M.N. Al-Ameen and M.D.R. Hasan, “The Mechanisms to Decide on Caching a Packet on Its Way of Transmission to a Faulty Node in Wireless Sensor Networks based on the Analytical Models and Mathematical Evaluations,” International Conference on Sensing Technology, pp. 336-341, 2008.

[8] D. Nikos, K. Dimitrios and M. Yannis, “Cooperative Caching in Wireless Multimedia Sensor Networks,” Mobile Networks and Applications, Vol. 13, No. 3-4, pp. 337-356, 2008.

[9] T.P. Sharma, R.C. Joshi and Manoj Misra, “Cooperative Caching for Homogeneous Wireless Sensor Networks,” International Journal of Communication Networks and Distributed Systems (IJCNDS), Vol. 2, No. 4, 2009.

[10] N. Dimokas, D. Katsaros, L. Tassiulas and Y. Manolopoulos, “High Performance, Low Complexity Cooperative Caching for Wireless Sensor Networks,” Springer International Journal of Wireless Networks, Vol. 17, No. 3, pp. 717-737, 2011.

[11] Amir Shiri, Shahram Babaie and Javad Hasan-zadeh, “New Active Caching Method to Guarantee Desired Communication Reliability in Wireless Sensor Networks,” Journal of Basic and Applied Scientific Research, Vol. 2, No. 5, pp. 4880-4885, 2012.

[12] Xiaorong Zhu, Lianfeng Shen, and Tak Shing Peter Yum, “Hausdorff Clustering and Minimum Energy Routing for Wireless Sensor Networks,” IEEE Transaction on Vehicular Technology, Vol. 58, No. 2, pp. 990-997, Feb 2009.

DNA Microarray: Identification of Biomarkers to Detect HCV Infected With Hepatocellular Carcinoma by the Analysis of Integrated Data

Salwa Eid, Aliaa Youssif, Samar Kassim

Abstract— Hepatocellular Carcinoma (HCC) is the one of leading causes of cancer related deaths worldwide. In most cases, the patients are first infected with Hepatitis C virus (HCV) which then progresses to HCC. HCC is usually diagnosed in its advanced stages and is more difficult to treat at this stage. Early diagnosis increases survival rate as treatment options are available for early stages. Therefore, accurate biomarkers of early HCC diagnosis are needed. DNA microarray technology has been widely used in cancer research. Scientists study DNA microarray gene expression data to identify cancer gene signatures which helps in early cancer diagnosis and prognosis. Most studies are done on single data sets and the biomarkers are only fit to work with these data sets. When tested on any other data sets, classification is poor. In this paper, we combined four different data sets of liver tissue samples (101 HCV-cirrhotic tissues and 57 HCV-cirrhotic tissues from patients with HCC). Differently expressed genes were studied by use of high-density oligonucleotide arrays. We extracted the most informative features using LASSO regression and Random Forest. Then applied different classifiers to distinguish HCV samples from HCV-HCC related samples using the genes selected.

.

Keywords— DNA microarray, HCV, HCC, feature selection, classifiers, integrative analysis

I. Introduction

It was estimated in 2002, liver cancer is the sixth leading cancer type, with 62,6162 cases, and is the third leading cause of cancer death, with an estimated 58,321 deaths in that year [1]. In Egypt, HCC was reported to count for about 4.7% of infected patients worldwide [2]. HCV and HBV are the main risk factors for the development of HCC([3]-[5]). Egypt has the highest prevalence of HCV worldwide and up to 90% of HCC infected Egyptian patients was caused by HCV[6]. HCC's only curative treatments are surgical resection and liver transplant ([7]-[9]). These treatments are not applicable to candidates in the late stages of HCC. The earlier HCC is detected followed by an appropriate treatment can reduce the number of deaths caused by tumours [10]. Serum α-fetoprotein determination and ultrasongraphy are used for HCC diagnosis. Although serum α- fetoprotein is cheaper, serum α-fetoprotien determination is not always a biomarker for HCC especially anything related to HCC-HCV. In a series of 606 HCC patients, normal serum α-fetoprotein was observed in 40.4% of patients with small HCC tumours, in 24.1% of patients with tumours 2 to 3 cm in diameter, and in 27.5% of patients with 3 to 5 cm tumours [11]. As for ultrasonography has been described as highly user dependent[12]. Therefore improved biomarkers for early and accurate detection of HCC are needed. Several studies have been done on HCV and HCV-HCC related samples using DNA microarray and gene sets were identified that could be useful as diagnostic tools. Most studies were assessed on single datasets via cross validation or tested on a small test set. It's been noticed among the research done, that each study conducted using DNA microarray on HCV and HCV-HCC related resulted in a different gene signature and there was no overlap in the genes reported informative or just a few of them. These results make it difficult to identify the most predictive genes for detecting HCC in HCV patients. This is due to the differences in the array platforms, experiments, experiments' condition and the underling biological heterogeneity of the disease. Therefore validation on large combined data set is still missing. The present study combined 4 datasets and has focused on finding biomarkers that can be used for the early detection of HCC caused by HCV. We investigated whether gene classifiers derived from two datasets using different array platforms could be independently validated by application to other datasets or not.

In the twentieth century, scientists used to study a single or a few genes at a time by southern blotting and northern blotting. Using such techniques, to study the human genes which are around 20, 000 genes require a lot of time and is quite hard. Therefore DNA microarray technology has evolved. It enables researchers to analyze thousands of genes expressions simultaneously. DNA microarray is used in a lot of fields such as gene discovery, disease diagnosis drug discovery and toxicological research. An important application of DNA microarray is the identification of genes that play significant roles in human carcinogenesis. Another application is prediction of a categorical class based on the expression profile of the patient.

In this paper we concentrated on finding new candidate genes that would classify between two classes HCV and HCV-HCC related samples. When creating a classifier two aspects are important: selecting the features, that are informative and choosing a classifier that performs well. In section 2, we talked briefly about data integration which is followed by a description of the existing algorithms in section 3. In section 4, we explain the work flow and propose a novel ensemble feature classifier. Experiment and results are in section 5 and section 6, respectively. Conclusion and suggested future work are then presented in section 7.

II. Data Integration

Due to the differences in research results on individual datasets, analysis of integrated datasets is needed. The integration of multiple datasets promises to yield more reliable and accurate results. The probability of the biomarkers being over fitted to a single dataset will be eliminated as the research will consist of multiple datasets. There are two methods to combine inter-study microarray data at different levels([13]-[16]). The first method is meta-analysis, which combines results from individual data sets to avoid the direct comparison of gene expression values and to increase the power of identifying significantly expressed genes among them. The second method is direct integration of expression values after specific data transformation and normalization on individual data sets. This method is done in two steps. First, a list of common genes of the multiple different microarray platforms are extracted based on cross-referencing the annotation of each probeset represented on the microarrays. The cross-referencing of expression data is done using formatted R-packages accessed online. Next, for each individual data set, numerically comparable quantities are derived from the expression values of the common genes by applying specific data transformation and normalization methods. Then new gene expression values from each datasets are combined to increase sample size and then the analysis is done to the merged data.

III. Constructing classifiers

There are two steps to building a prediction model to differentiate between two conditions which are: gene reduction and choosing the most appropriate classifiers. It is considered as an optimization problem, selecting the best features with the best classifier which would give the minimum prediction error.

A. Feature Extraction

Logistic regression models are commonly used when working with HCV and HCV-HCC classes but shouldn't be used when then number of predictor variables (p) exceeds the sample size (n) [17]. Three linear regression algorithms Least Angle Regression (LAR), Least Absolute Shrinkage Operator (LASSO) and Average Linear Regression (ALM) were evaluated in the prediction of classes on high dimensional gene expression data by Yingdong Zhao [18]. It was demonstrated that LAR and LASSO perform quite well and in a similar manner when used on data without noise and better than ALM but LASSO performed best on data with noise. In such cases, penalized methods are thought to give better results ([19], [20]). The LASSO is a penalized method for estimating a logistic regression model when p>n [21]. The LASSO model is estimated using maximum likelihood [21]. Regression coefficients are calculated for all genes to minimize a weighted average of mean squared prediction error for the training set plus the sum of absolute values of all regression coefficients. It tends to assign zero coefficients to genes that are less informative. Then weighting factor is optimized by cross-validation [21]. LASSO algorithms attempt to avoid the over-fitting characteristic of least-squares linear regression when the number of variables is large compared to the number of cases [22].

The Random Forest is one of the methods used in feature selection in tumor problems. It is a classification algorithm that directly provide measures of variable importance that are of great interest for gene selection. Random forest was developed by Leo Breiman[23]. Random forest returns several measures of variable importance. One of the important measures is based on the reduction of classification accuracy, when values of a variable in a node of a tree are tested randomly [24]. To select genes, random forests are fit iteratively and at each iteration the new forest is build after discarding those features with the smallest variable importance.

B. Classifiers

Hedenfalk used compound covariate classifier , to predict the *BRCA1* and *BRCA2* mutation status of breast cancer specimens [25]. The Compound Covariate Predictor is a weighted linear combination of log-ratios for genes that are univariately significant at a specified level. A two-sample t-test is performed. That is differentially expressed genes with log-expression ratios that best discriminates between the two classes and meets the specified level α, are selected. Then a single compound covariate is constructed using the differentially expressed genes for class prediction. The two-sample t statistic of each differentially expressed gene acts as its weight in the compound covariate. Thus, the value of the compound covariate for sample *i* is

$$C_i = \sum_j t_j \quad x_{ij} \qquad (1)$$

where t_j is the *t*-statistic for the two group comparison of labels with respect to gene *j*, x_{ij} is the log-ratio measured in sample *i* for gene *j* and the sum is over all differentially expressed genes. After the value of the compound covariate is computed for each sample in the training set, a classification threshold is calculated C_t.

$$C_t=(C_1+C_2) / 2 \qquad (2)$$

,where *C*1 and *C*2 are the mean values of the compound covariate for samples in the training set with class label 1 and class label 2, respectively. C_t is the midpoint of the means of the two classes. A new specimen is predicted to be of class 1 if its compound covariate is closer to C_1 and to be of class 2 if its value is closer to C_2[26].

Diagonal Linear Discriminant Analysis is similar to Compound Covariate Predictor. It has the same prediction rule but with equal prior probabilities for the two classes [27]. It also ignores the correlations among the genes to avoid over-fitting the data. Dudoit. reported that performance of simple classifiers such as linear discriminant analysis and the nearest neighbour performed well as much more sophisticated methods such as aggregated classification trees[28].

The *K*-Nearest Neighbor Predictor depends on the majority voting of the *k*-nearest neighbors. A sample is assigned to class if the expression profile of it, gets the most votes to the k-training samples in that class. Euclidean distance is used to calculate the distance metric[29].

Nearest Centroid predictor calculates the centroid for each class which is the average gene expression for each gene in each class divided by the within-class standard deviation for that gene[30]. When predicting a new sample, the gene expression of that sample is compared to all of the class centroids. The class whose centroid, that it is closest to, in squared distance, is the predicted class for that new sample.

Prediction Analysis of Microarray(PAM) is another method that uses the shrunken centroid algorithm developed by Tibshirani[31]. It is a modified method to Nearest Centroid. It shrinks each of the class centroids toward the overall centroid for all classes by a certain amount, threshold. This shrinkage consists of moving the centroid towards the threshold which is set by the user. Then the new sample is classified by the usual nearest centroid rule, but using the shrunken class centroids.

Support Vector Machine(SVM) is another effective prediction method. It's a machine learning technique[32]. We used the SVMS with linear kernel functions. It is a linear function of the log-intensities that best differentiates the data with respect to penalty costs on the number of samples misclassified.

Not only Random Forest can be used for feature selection but also for classification. It is an ensemble learner constructed from many classifiers. It combines the results of all classifiers used. It is robust against over fitting and usually performs better than other classifiers. It consists of multiple random trees classifiers that all vote on classification for a given set of inputs. An input is classified to class on which gets the most votes. In a study by Valeria, Random Forest was used for classification and feature selection[33]. Fifteen probesets were identified to classify between HCV and HCV-HCC related samples.

C. *Cross Validation*

Cross-Validation is a way of assessing the class prediction model and the feature selection model. With *K*-fold cross validation or leave-one-out cross-validation (LOOCV), the samples are randomly partitioned into *K* equal size groups $S_1,S_2,\ldots,S_K$. One of the *K* subsets is omitted and a model is build including the determination of which genes are univariately significant on the remaining training set which consists of the other *K-1* subsets. Then using that gene list, a multivariate predictor is constructed and applied to the sample that was omitted. This is repeated, omitting all of the samples one at a time. Cross-validated misclassification rate is computed using the number of samples misclassified during this process. Due to the large number of features, it is important to use cross-validation or some similar method to determine the accuracy of the model.

IV. WORK FLOW

In the current study, we aimed to generate a large enough dataset to create a highly genereralizeable set of minimum signatures that would be able to differentiate between HCV and HCV-HCC classes. We integrated datasets and used hybrid feature extraction classifier. The creation of predictive signatures consists of two steps: eliminating the less informative genes and selecting a predictive model with high performance in correctly predicting an HCV or an HCV-HCC related sample. Each dataset was read and normalized independently. Then the datasets were merged together. The next step is feature extraction. Instead of using a single feature extraction method, we used two methods and extracted the features in two stages. Using the LASSO regression model we determined the differentially expressed genes between the two classes. Then we took those differentially expressed genes and used Random Forest to eliminate the number of genes ending with genes that have the highest discriminatory ability to correctly classify a sample. Cross validation was performed in each stage to validate the differentially expressed genes obtained. We then used class prediction methods to determine whether the genes selected by the feature extraction phase accurately classified the HCV and HCV-HCC related samples. Cross validation was also performed here. The algorithm is summarized in fig.1.

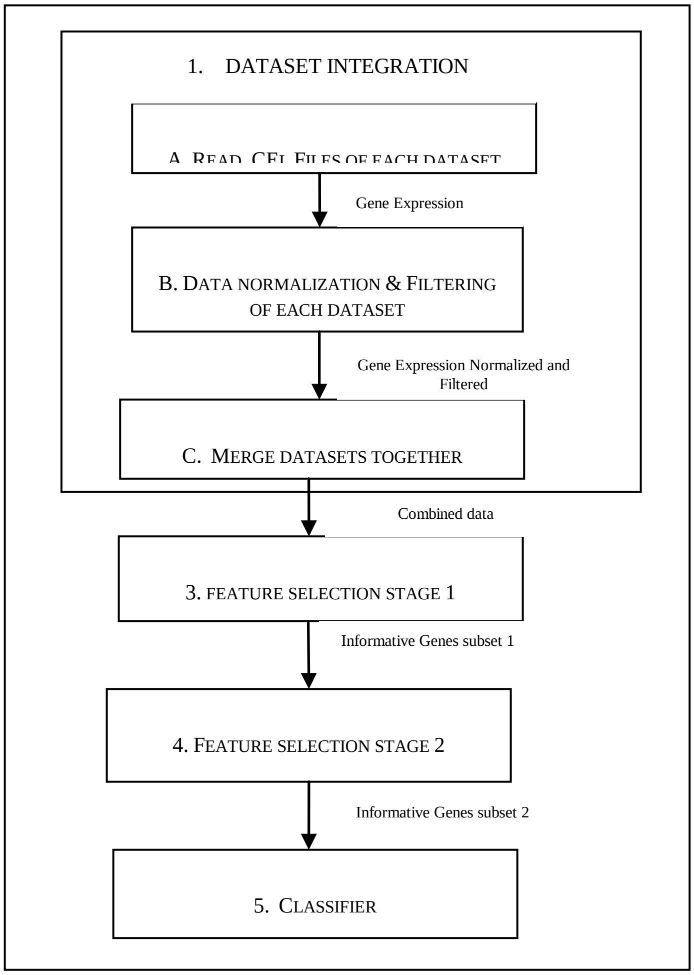

1. DATASET INTEGRATION
A. READ CEL FILES OF EACH DATASET
Gene Expression
B. DATA NORMALIZATION & FILTERING OF EACH DATASET
Gene Expression Normalized and Filtered
C. MERGE DATASETS TOGETHER
Combined data
3. FEATURE SELECTION STAGE 1
Informative Genes subset 1
4. FEATURE SELECTION STAGE 2
Informative Genes subset 2
5. CLASSIFIER

Figure 1: Work Flow

V. Experiment

A. *Figures and Tables*

A summary of the four datasets are given in table. 1. The four datasets were obtained from the Gene Expression Omnibus (GEO) database publically available over the internet[34]. Each obtained from different institutes. Raw data (.cel files) from each dataset were imported into BRB array tools[35]. The HCV samples were the only ones retrieved from the first dataset. As for the second data set, we only used the HCV and early stage and very early stage HCC samples. For the third dataset, we read the whole data. It represented HCV and HCV- HCC related samples only. As for the fourth the data set we retrieved only the HCV-HCC related samples.

Table 1 Data Sets Summary

GEO ID #	**D1** **GSE14323**	**D2** **GSE6764**	**D3** **GSE17967**	**D4** **GSE19665**
Institution	Virginia Common-wealth University	Mount Sinai School of Medicine	Virginia Common-wealth University	University of Tokyo
Pubmed ID *	19098997	17393520	19861515	20345479
Chip type	Affymetrix Human Genome U133A Array	Affymetrix Human Genome U133 Plus 2.0 Array	Affymetrix Human Genome U133A 2.0 Array	Affymetrix Human Genome U133 Plus 2.0 Array
Number of Samples	58	31	63	5

B. *Normalization and Filtering*

Each data-set was normalized independently using the GC-RMA algorithm ([36],[37]). Arrays in each dataset are normalized to a reference array. The reference array is the array whose median log-intensity value is the median over all median log-intensity values for the set of the array. GC-RMA methods adjust for background intensities that include optical noise and non-specific binding. A background correction on the perfect match(PM) is done, followed by quantile normalization. Then the probe set summaries information are obtained using Tukey's median polish algorithm[38]. Each dataset was filtered so that affymetrix control-probe set genes were excluded. The 22,215 common RMA probe-sets expression summaries across the three datasets were merged manually.

C. *Simulation*

Six experiments were done. Details of the six experiments are show in table. 2. The first three are individual analysis and other three are integrative analysis. We applied the LASSO to the training data in each experiment. We tried different values for K for the cross validation and ended up choosing k=10 as it gave us the lowest prediction error. Each time, 10% of the samples are omitted; the model is built using the remaining 90% of the samples. The prediction errors are recorded for the samples withheld. This is done 10 times, omitting each of the 10 subsets one at a time and the errors for the samples in each subset are obtained and totalled into an overall error. We avoided the leave-one-out cross-validation to avoid an over fitted model over the training data. The predicted class was HCV-HCC related sample if the fitted probability was greater or equal to 0.5, otherwise an HCV sample. As we can see in table 2, Exp 4 gave us the best results for both the cross-validation and the correct classification of the new samples. The genes included in the LASSO model done on Exp4, along with their coefficients and a percentage of CV support which provides information on the stability of gene selection across loops of the cross-validation are shown in table. 3.

Table 2. LASSO Model Results for all Experiments

Exp	**Training**	**Testing**	**Cross-validation**	**Correct Classification**	**Number of genes**
Exp1	D1	D2, D3, D4	88%	70%	14
Exp 2	D2	D1, D3, D4	83%	73%	19
Exp 3	D3	D1, D2, D4	100%	74%	11
Exp 4	D2,D3	D1, D4	100%	95%	38
Exp 5	D1, D2	D3, D4	88%	86%	8
Exp 6	D1, D3	D2, D4	87%	94%	6

Therefore we continued to work with Exp4. We applied Random Forest to the 38 genes obtained from lasso model performed on Exp4. The random forest minimized the number of genes to 25. Using these features selected, we constructed the following classifiers: PAM, Compound covariate predictor, Diagonal linear discriminate analysis, nearest neighbour predictor for K=1 and K=3, nearest centroid predictor, support vector machine predictor and random forest.

VI. Results

A. *Comparison of integrative analysis with individual analysis for feature selection*

As we can see the integrative analysis for the three experiments gave higher results than the other three experiments. Sensitivity obtained for each dataset was 38.5%, 40%, 13.5%, 91%, 78% and 95% for Exp1, Exp2, Exp3, Exp4, Exp5 and Exp6 respectively. As for specificity obtained was 90%, 98%, 100% ,98%, 100% and 94% for Exp1, Exp2, Exp3, Exp4, Exp5 and Exp6 respectively. We have also calculated the area under the curve(AUC) for the receiver operating characteristic(ROC) curve for the 6 experiments as an evaluation. The ROC curve is a graphical plot of true positive(sensitivity) against false positive(1-specificity). It illustrates the quality of performance of a classifier. The greater the AUC, meaning the higher the classifier performs. An AUC of 1 means perfect classifier. The following are the values of AUC: 0.642, 0.691, 0.566 and 0.942, 0.889 and 0.912 for Exp1, Exp2, Exp3, Exp4, Exp5 and Exp6 respectively. As observed, Exp4 has the highest AUC reflecting that it's the best model.

Table 3: Genes Obtained from LASSO Model (Exp4)

Gene Symbol	**Coefficient**	**% CV Support**
BGN	0.00115	50
ALDH3A2	-0.04653	20
S100A13	0.47828	90
CD97	0.526	90
S100A8	0.07493	60
C1QB	0.13903	40
SSX2IP	-0.27354	90
RAB4A	-0.09397	10
DPP4	-0.34647	100
KIAA0467	-0.27724	30
PROM1	0.60033	100
GALC	0.13688	50
ELMO1	-0.10681	40
CACNA2D2	-0.18859	50
EEA1	-0.28478	90
NA	0.60932	100
PTPN4	-0.26679	60
ZNF264	-0.35938	100
AKR1B10	-0.24597	100
RBPMS	1.29121	100
RPL39	-2.63319	100
NA	-0.2691	80
SNRK	-0.21413	30
RBPMS	0.22174	90
SLC19A1	-0.03816	20
HAUS3	-0.16743	50
FKBP15	0.06006	20
RNF115	-0.08253	70
ZEB1	-0.3381	90
CALR	0.75398	80
C14orf147	-0.01653	20
CALR	0.06269	70
MGEA5	-1.31245	90
HECTD3	0.07716	10
NA	-0.22339	40
GLTP	0.12276	40
ZNF432	-0.117	30
ACACB	0.11383	40

B. Comparison of Different Classifiers using Features selected

After eliminating the number of genes using Random Forest, we used different classifiers for prediction and compared the results. It shows that the K-nearest neighbor classifier with k set to 1 and also k set to 3 gave us the lowest misclassification error of 4.5% when compared with the rest of classifiers. We plotted the ROC curves for the all the predictors as shown in fig. 2 and the AUC was calculated. The values of AUC obtained were 0.878, 0.923, 0.923, 0.905, 0.999,0.999, 0.824 and 0.961 for PAM, Random Forest, Linear Discriminant Analysis, Compound Covariate, 1-Nearest neighbor, 3-Nearest neighbor, Nearest Centroid and Support Vector Machine respectively. As observed, 1-Nearest neighbor and 3-Nearest neighbor tied and gave us the highest AUC.

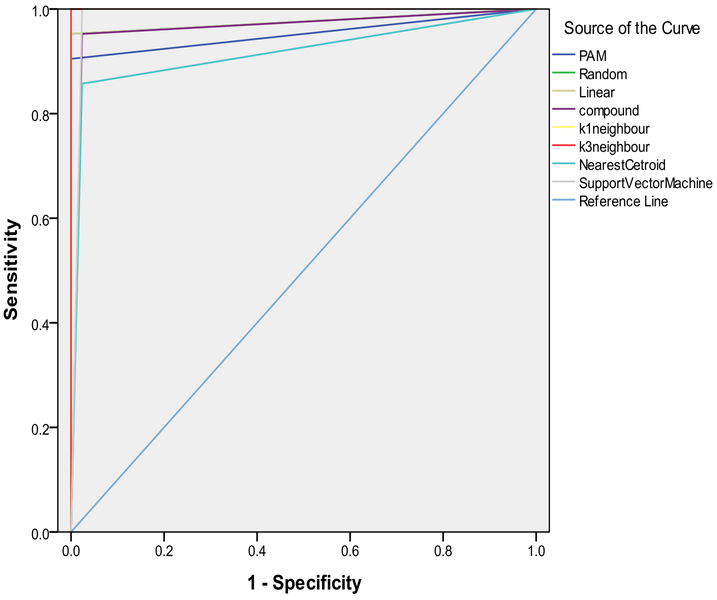


Figure 2: ROC curve for all Classifiers

VII. Conclusion and Future Work

Our aim was to develop new valid biomarkers for HCV cirrhotic patients with or without HCC. Most studies were done on single datasets and the resulting classifier and features selected were over fitted to these datasets. Our study combined different datasets from different hospitals and an ensemble feature extraction classifier was constructed. We filtered the less informative genes on two stages. We reduced the 22,215 genes to 38 genes during the first stage using LASSO regression model and then reduced the 38 genes to 25 during the second stage using Random Forest. We evaluated the signatures by applying the different classifiers to validate the genes selected. The average for the miss classification error for all the classifiers used using the 25 features was 6%. We assessed the classifiers using the ROC curve and the AUC. The AUC showed that the K-nearest neighbour with K set to 1 and K set to 3 were the best classifiers used with a AUC value of 0.999. The other classifiers also performed well, indicating that the genes selected are real candidate biomarkers for HCV-HCC patients. Most studies reached biomarkers that perform well on certain datasets and using certain classifiers but when tested on different datasets and using different classifiers performed poorly. The signatures identified in this research are general and are not over fitted for one single dataset. The 25 features, a subset of the 38 genes with the combination of the classifiers used in this paper and other classifiers could be tested on other datasets. These genes could be tested clinically and turned into real diagnostic biomarkers that could be used. It is of great importance to identify signatures that could identify the presence of HCC in cirrhotic tissues which would help in early diagnosis of cancer.

References

[1] Parkin DM, Bray F, Ferlay J, Pisani P.(2001)Estimating the world cancer burden*: Globocan 2000. Int. J. Cancer* 94:153-156

[2] Rahman El-Zayadi A, Abaza H, Shawky S, Mohamed MK, Selim OE, Badran HM: Prevalence and epidemiological features of hepatocellular carcinoma in a Egypt-a single center experience. *Hepatol Res* 2001, 19(2):170-179

[3] Block TM, Mehta AS, Fimmel CJ, Jordan R:Molecular viral oncology of hepatocellular carcinoma. *Oncogene* 2003, 22:5093-5107.

[4] Buendia MA: Hepatitis B viruses and cancerogenesis. *Biomed Pharmacother* 1998, 52:34-43

[5] Colombo M: The role of hepatitis C virus in hepatocellular carcinoma. *Recent Results Cancer Res* 1998, 154:337-334

[6] Goldman R, Ressom HW, Abdelhamid M, et al. Candidate markers for the detection of hepatocellular carcinoma in low molecular weight fraction of serum. *Carcinogenesis*. 2007; 28(10):2149-2153

[7] Marsh JW, Dvorchik I. (2003) Liver organ allocation for hepatocellular carcinoma: are we sure? *Liver Transpl*. 9:693-696.

[8] Fisher RA, et al.(2007) Is hepatic transplantation justified for primary liver cancer? J Surg. *Oncol*. 95:674-679

[9] Barnett CC Jr, Curley SA. (2001) Ablative techniques for heptocellular carcinoma. Semin. *Oncol*. 28:487-496.

[10] Llovet JM, Burroughs A, Bruix J. Hepatocellular carcinoma. *Lancet* 2003; 362:1907-1917.

[11] Nomura F, Ohnishi K, Tanabe Y. Clinical features and prognosis of hepatocellular carcinoma with reference to serum alpha-fetoprotein levels. Analysis of 606 patients. *Cancer* 1989; 64:1700-1707

[12] Beale G, Chattopadhyay D, Gray J, et al. AFP, PIVKAII, GP3, SCCA-1 and follisatin as surveillance biomarkers for hepatocellular cancer in non-alcoholic and alcoholic fatty liver disease. *BMC Cancer* 2008:8:200

[13] Ghosh, D., et al., Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Functional & Integrative Genomics*, 2003. **3**:p. 180-188.

[14] Rhodes, D.R., et al., Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer. *Cancer Res*, 2002. **62**(15): p. 4427-4433.

[15] Choi, J.K., et al., Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 2003. **19**(90001): p. 84-90

[16] Stevens, J. and R.W. Doerge, Combining Affymetrix microarray results. *BMC Bioinformatics*, 2005. **6**(1): p. 57..

[17] Kellie J.Archer, Valeria R. Mas, Krystle David, Daniel G. Maluf, Karen Bornstein, Robert A. Fisher, Identifying Genes for Establishing a Multigenetic Test for Hepatocellular Carcinoma Surveillance in Hepatitis C Virus-Positive Cirrhotic Patient. *Cancer Epidemiol Biomarkers* 2009:11:2929-2932

[18] Y. Zhao, R. Simon, Development and Validation of Predictive Indices for a Continuous Outcome Using Gene Expression Profile. *Cancer Informatics*., vol. 9, pp.105-114, 2010

[19] Gui J, Li H. Penalized Cox regression analysis in high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* 2005:21:3001-3008

[20] Ma S, Song X, Huang J. Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics* 2007;8:60.

[21] Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med* 1997;16:385:395

[22] Efron B, Hastie T, Johnstone I and Tibshirani R (2004). *Least Angle Regression.Annals of Statistics 32* (2): pp. 407–499.

[23] Breiman L: *Random forests. Machine Learning* 2001, 45:5-32.

[24] Bureau A, Dupuis J, Hayward B, Falls K, Van Eerdewegh P: Mapping complex traits using Random Forests. *BMC Genet* 2003,4(Suppl 1):S64.

[25] Hedenfal, I., Dugga, D., Chen Y., Radmache, M., Bittner M., Simon R., Meltzer P., Gusterson B., Esteller M., Gene-expression pro. les in hereditary breast cancer., *N. Engl. J. Med*,344, pp. 539–548,2002

[26] Radmacher MD, McShane LM and Simon R. A paradigm for class prediction using gene expression profiles. *J Comput Biol* 2002;9:505–12.

[27] Morrison, D. F, Multivariate statistical methods ,*McGraw-Hill*: New York,1967,pp. 130-133.

[28] Raffeld, M.,S. Dudoit, J. Fridlyand, and P. Speed, Comparison of discrimination methods for classification of tumors using gene expression data, *J.Amer. Statist. Assoc*., vol. 97, 2002, pp. 77–87.

[29] Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method.Bioinformatics 2001;17:1131–42.

[30] Dabney,A.R. Classification of microarrays to nearest centroids. Bioinformatics,2005, 21, 4148–4154.

[31] Furey TS. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16(10):906–14.

[32] Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2002) *Proc. Natl. Acad.Sci. USA* **99,** 6567–6572.

[33] Mas, V. R., D. G. Maluf, K. J. Archer, K. Yanek, X. Kong, L. Kulik, C. E.Freise, K. M. Olthoff, R. M. Ghobrial, P. McIver, and R. Fisher. Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virusinduced hepatocellular carcinoma. *Mol. Med.2009*: 15:85–94.

[34] U.S. National Library of Medicine, National Center for Biotechnology Information, GEO Datasets. [Online]. Available: http://www.ncbi.nlm.nih.gov/gds/ [Accessed: 30 Sept. 2011]

[35] Richard Simon, National Cancer Institute, Biometric Research Branch, BRB Array tools, 2002. [Online]. Available: http://linus.nci.nih.gov/BRB-ArrayTools.html#content [Accessed: 30 Sept. 2011]

[36] Richard, S, Lam A, Li MC, Ngan M, Menenzes S, et al. (2007) Analysis of Gene Expression Data Using BRB-Array Tools. *Cancer Inform* 3: 11–17.

[37] Wu Z, Irizarry RA (2004) Preprocessing of oligonucleotide array data. *Nat Biotechnol* 22: 656–658; author reply 658.

[38] RAIrizarry, et. Al. “Summaries of Affymetrix Gene Chip probe level data. *Nucleic Acids, Research*, 2003, vol.31, No.4.

Implementation for Model of Object Oriented Class Cohesion Metric -MCCM

Tejdeda Alhussen Alhadi
tt_hussen@yahoo.com

Dr. Omer Saleh
immer.jomah@gmail.com

Xavier Patrick Kishore
patrick.kishore@gmail.com

Sagaya Aurelia
sagaya.aurelia@gmail.com

Department of Computer Science
Faculty of Education
Beniwalid, Libya

***Abstract*—Class cohesion should not exclusively be based on common instance variables usage criteria. Method Connectivity Cohesion Metric (MCCM) uses both direct and indirect method relations (attributes usage criterion and methods invocation criterion) in its calculations [1]. This paper presents a tool to measure MCCM (Method Connectivity Cohesion Metric) that measures the cohesion of classes coded by Java programming language. The major motivation is to carry out this study that computes the class cohesion and comparison between MCCM metric and LCC.**

Keywords*—*MCCM; Cohesion; Method Invocation; Attribute Usage

I. INTRODUCTION

There are lots of metrics for coupling and cohesion in the literature, for the estimation of class cohesion is based on different relationships that may exist between its methods. It takes into account several ways of capturing the functional cohesion of the class, by focusing on Connectivity between methods [1]. The MCCM tool measures the quality of the entrance program by the user, Then entrance program is analyzed to classes, these classes are measured its cohesion and coupling of each class separately by the used metric (MCCM). Finally, (MCCM) tool presents the results of cohesion and coupling of each class, to make the programmed able to modify any class has non-satisfactory results in order to reach the highest quality.

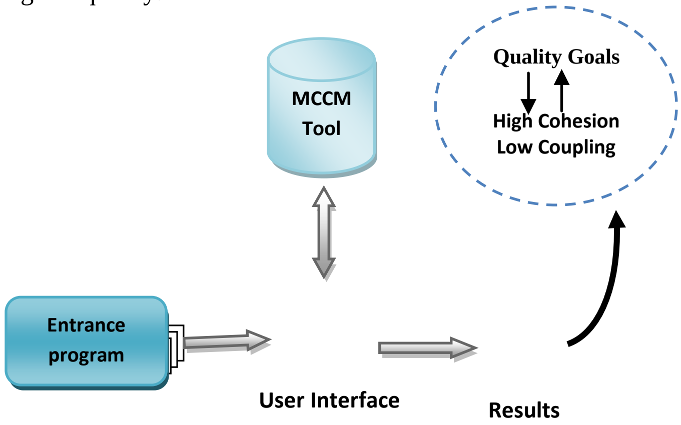


Fig.1. Overview of MCCM Implementation

II. ENVIRONMENT

The Java runtime environment (JRE) is required in order to run (MCCM.jar). We have also used Eclipse (SDK 3.1) which is a kind of universal tool platform-an open extensible IDE for anything and nothing in particular. It provides a feature rich development environment that allows the developer to efficiently create tools that integrate seamlessly in Eclipse platform [2]. The jar file is directly executable by Eclipse SDK program.

A. *MCCM Metric Architecture*

MCCM metric is a software metrics tool that measures the structural properties of java code and computes a number of software measures that include cohesion and coupling.

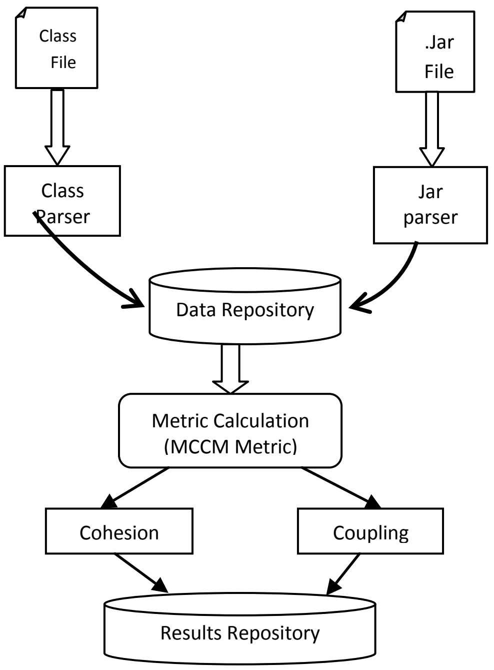


Fig.2. MCCM Metric Architecture

As shown in Fig.2, object-oriented systems are parsed to the tool in order to collect the data that can be used in computing the various software metrics supported by the tool. The data collected are stored in central data repository and results in result repository.

B.Flow Chart of (MCCM) Tool

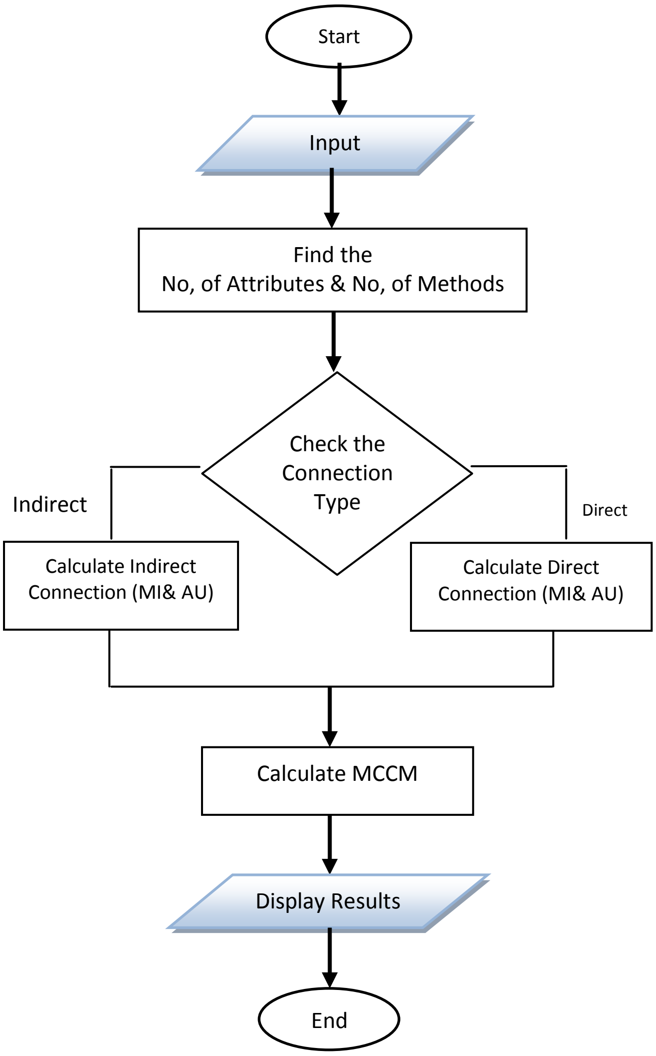
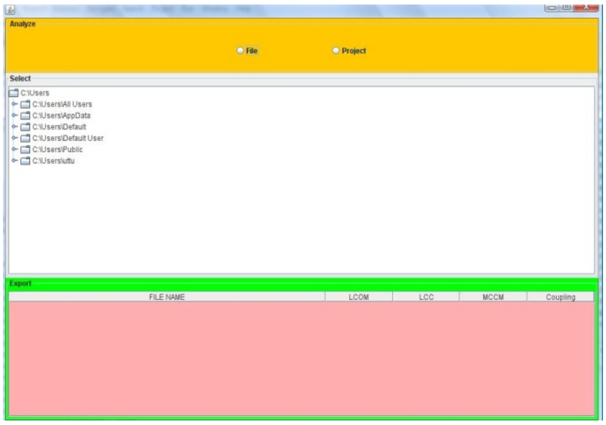


Fig.3. Flow Chart of (MCCM) Tool

III. Using MCCM Measurement Tool

This program has three main sections:
A. Input Section
B. Analyze Section
C. Output Section



○ File ○ Project

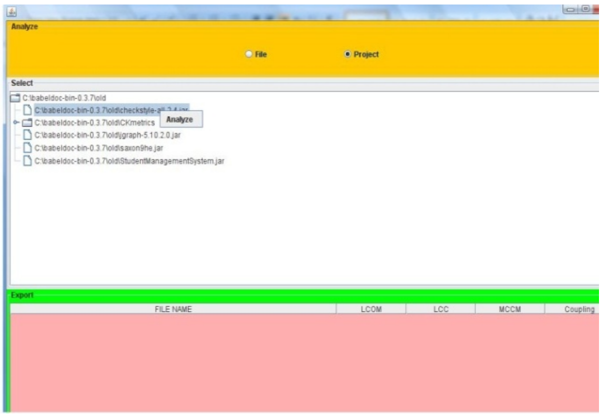
```
C:\Users
C:\Users\All Users
C:\Users\AppData
C:\Users\Default
C:\Users\Default User
C:\Users\Public
C:\Users\uttu
```

FILE NAME	LCOM	LCC	MCCM	Coupling

Fig(4) main page of MCCM software.

A. Analyzing the file

file chosen in the first step will be listed. To analyze which is the second section of this program, right click on file (.class or .jar) and select Analyze.



○ File ● Project

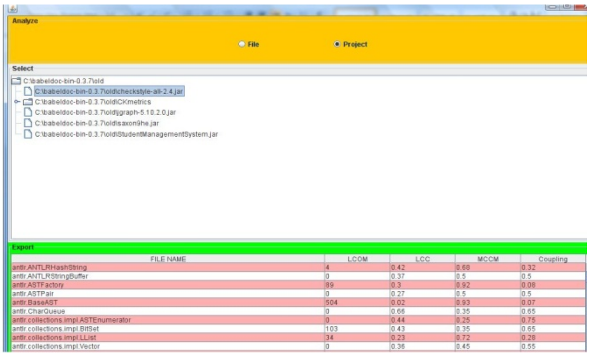
```
C:\babeldoc-bin-0.3.7\old
C:\babeldoc-bin-0.3.7\old\checkstyle-all-2.4.jar
C:\babeldoc-bin-0.3.7\old\CKmetrics
C:\babeldoc-bin-0.3.7\old\jgraph-5.10.2.0.jar
C:\babeldoc-bin-0.3.7\old\saxon9he.jar
C:\babeldoc-bin-0.3.7\old\StudentManagementSystem.jar
```

FILE NAME	LCOM	LCC	MCCM	Coupling

Fig.5. Screen Capture of Analysis the file

B.Examining the results

The results will be listed in the output section, the third section of this program.



○ File ● Project

```
C:\babeldoc-bin-0.3.7\old
C:\babeldoc-bin-0.3.7\old\checkstyle-all-2.4.jar
C:\babeldoc-bin-0.3.7\old\CKmetrics
C:\babeldoc-bin-0.3.7\old\jgraph-5.10.2.0.jar
C:\babeldoc-bin-0.3.7\old\saxon9he.jar
C:\babeldoc-bin-0.3.7\old\StudentManagementSystem.jar
```

FILE NAME	LCOM	LCC	MCCM	Coupling
antlr.ANTLRHashString	4	0.42	0.68	0.32
antlr.ANTLRStringBuffer	0	0.37	0.5	0.5
antlr.ASTFactory	89	0.3	0.92	0.08
antlr.ASTPair	0	0.27	0.5	0.5
antlr.BaseAST	504	0.02	0.93	0.07
antlr.CharQueue	0	0.66	0.35	0.65
antlr.collections.impl.ASTEnumerator	0	0.44	0.25	0.75
antlr.collections.impl.BitSet	103	0.43	0.35	0.65
antlr.collections.impl.LList	34	0.23	0.72	0.28
antlr.collections.impl.Vector	0	0.36	0.45	0.55

Fig.6.Screen Capture of Exporting the results

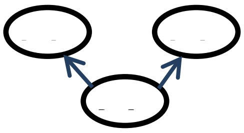
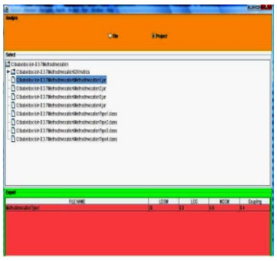
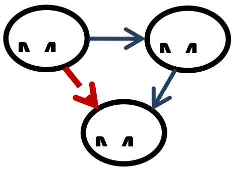
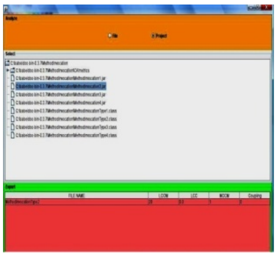
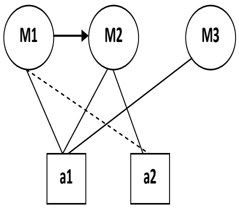
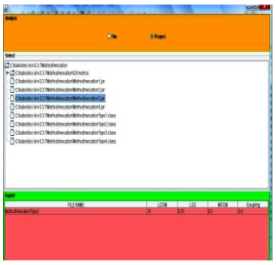
IV. The Evaluation

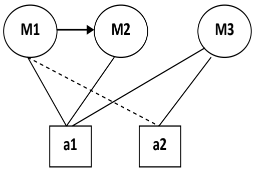
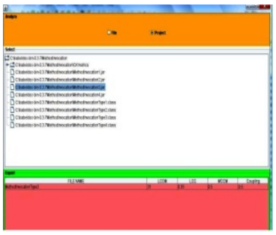
In order to complete the project and make it ready for the user it should undergo an evaluation process to confirm that MCCM tool fulfills the proposed objectives and demonstration of the software.

A. ConnectivityCriteria

In this section will calculate for each connectivity (Direct / Indirect) mathematical MCCM value [1] and the compare with MCCM tool value. The table (1) shows the results of this comparison.

Table 1 Connectivity Criteria

MCCM ***(Mathematical Calculation)***	***MCCM*** ***(Metric Output)***
 $\frac{[(0+0)+(1+0)+(1+0)]}{\left(\frac{3\ X\ (3-1)}{2}\right)+\left(\frac{3\ X\ (3-1)}{2}\ X\ 0\right)}=\frac{2}{3}$	
 $\frac{[(1+0)+(1+0)+(1+0)]}{\left(\frac{3\ X\ (3-1)}{2}\right)+\left(\frac{3\ X\ (3-1)}{2}\ X\ 0\right)}=1$	
 $\frac{[(1+2)+(0+1)+(0+1)]}{\left(\frac{3\times(3-1)}{2}\times 1\right)+\left(\frac{3\times(3-1)}{2}\times 2\right)}=\frac{5}{9}$	

B.Selected System

From a total of six projects, three projects were collected from www.projectsparadise.com, and other projects from SourceForge.net which are an open source websites that provides a centralized space where open source developers can control and manage open source software development [3] [4].

Table 2 Details of the Projects

Project	*No of Classes*	*No of methods*	*No of Attributes*
bluej-307	7	86	245
car_sales_system	11	102	323
LibraryManagementSystem	29	570	340
checkstyle-all-2.4	78	492	228
jgraph-5.10.2.0	50	750	340
Saxon9he	107	1252	876

C.Snapshot of Output of Projects Using MCCM Tool:

Table 3 Snapshot of Output

Project name	*Class name*	*LCOM*	*LCC*	*MCCM*	*COUPLING*
Bluej-307.jar	Org.apache.log4j.chainsaw.controlPanel$1	1	0.11	0.33	0.67
	Org.apache.log4j.chainsaw.controlPanel$2	1	0.32	0.33	0.67
	Org.apache.log4j.chainsaw.controlPanel$3	100	0.02	0.33	0.67
Car_sale_system	AboutDialog	4	0.08	0.88	0.12
	AddCarPanel	7	0.15	0.83	0.17
	Car	77	0.01	0.88	0.12
Library managment system	AddBooks	12	0.01	0.93	0.07
	AddMembers	3	0.04	0.88	0.12
	Books	102	0.0	0.92	0.08
Checkstyle.jar	Antlr.ANTLRHashString	4	0.42	0.68	0.32
	Antlr.ANTLRStringBuffer	0	0.37	0.5	0.5
	Antlr.ASTFactory	89	0.3	0.92	0.08
	Antlr.ASTPair	0	0.27	0.5	0.5
	Antlr.BaseAST	504	0.02	0.93	0.07
	Antlr.CHarQueue	0	0.66	0.35	0.65
Jgraph.jar	Org.jgraph.event.GraphSelectionEvent	0	0.41	0.25	0.75

	Org.jgraph.AbstractCellView	383	0.07	0.85	0.15
	Org.jgraph.graph.AttributeMap	325	0.03	1.0	0.0
	Org.jgraph.graph.BasicMarqueeHandler	88	0.07	0.75	0.25
	Org.jgraph.graph.ConnectionSet$Connection	9	0.11	0.66	0.34
	Org.jgraph.graph.ConnectionSet	30	0.46	0.5	0.5
Saxon9he.jar	Javax.xml.xquery.XQConstant	1	0.0	1.0	0.0
	Javax.xml.XQQueryException	2	0.1	0.71	0.29
	Javax.xml.xquery.XQStackTraceVariable	0	0.22	0.5	0.5
	Net.sf.saxon.dom.AttrOverNodeInfo	46	0.09	0.63	0.37
	Net.sf.saxon.dom.DocumentBuilderFactoryImpI	17	0.13	0.7	0.3
	Net.sf.saxon.dom.DocumentBuilderImpI	15	0.18	0.61	0.39

D.Results and Analysis

TABLE (4) RESULTS OF (CAR_SALES_SYSTEM.JAR)

Project name	***Class name***	***LCOM***	***LCC***	***MCCM***	***Coupling***
car_sales_system	AboutDialog	4	0.08	0.88	0.12
	AddCarpanel	7	0.15	0.83	0.17
	Car	77	0.01	0.88	0.12
	CarDrtailsComponents	21	0.05	0.85	0.15
	CarSalesSystem	130	0.02	0.95	0.05
	CarsCollection	0	0.23	0.68	0.32
	Manufacturer	3	0.19	0.58	0.42
	SearchByAgepanel	0	0.26	0.72	0.28
	SearchByOtherpanel	0	0.22	0.77	0.23
	ShowAllCarpanel	0	0.34	0.62	0.38
	Welcomepanel	0	0.41	0.69	0.31

In this section we present the analysis of the results using MCCM tool, and Comparison between LCOM metric, LCC metric and MCCM metric. Table (4) shows the results of metrics calculation under mentioned project using LCOM, LCC and our proposed metrics for class cohesion MCCM and we can interpret the results as follows:

- The cohesion value of the LCC metric in class (AboutDialog) equals (0.08), and the cohesion value of the MCCM metric equals (0.88). The cohesion value of the LCC metric in class (AddCarpanel) equals (0.15), and the cohesion value of the MCCM metric equals (0.83). The cohesion value of the LCC metric in class (SearchByAgepanel) equals (0.26), and the cohesion value of the MCCM metric equals (0.72). We can notice that the MCCM metric tool gives higher value more than LCC metric value.
- The cohesion value of the LCOM metric in class (CarsCollection) equals (0.0), and the cohesion value of the MCCM metric equals (0.68). The cohesion value of the LCOM metric in class (SearchByOtherpanel) equals (0.0), and the cohesion value of the MCCM metric equals (0.77). The cohesion value of the LCOM metric in class (ShowAllCarpanel) equals (0.0), and the cohesion value of the MCCM metric equals (0.62). Where LCOM metric gives zero value, MCCM metric gives values which are higher than zero. So not all the connection types between elements in a class are taken into account in LCOM metric.
- The coupling values of all classes are low, as we know the maximum value of coupling is one. For example: coupling of class (Car) equals (0.12), and coupling of class (Welcomepanel) equals (0.31).

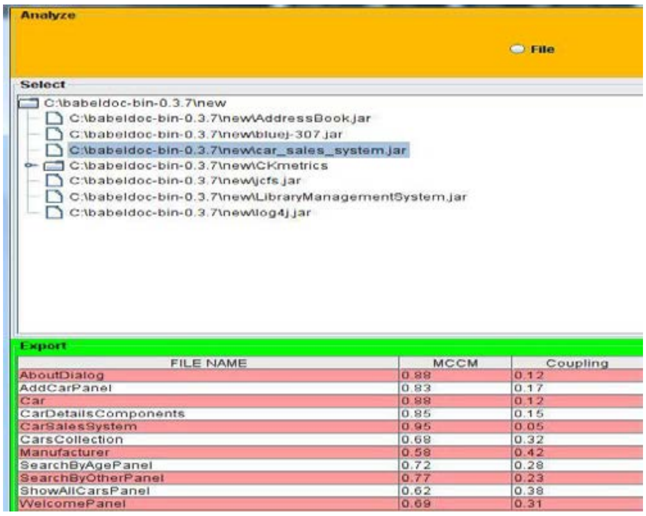
Metric	***Interaction Type***		***Interaction Mode***		***Method Interaction***	
	M→A	M→M	Direct	Indirect	M/ Invocation	A/ Sharing
LCOM	√		√			√
LCC		√	√	√		√
MCCM	√	√	√	√	√	√

Through this explanation we results that the above tables prove that MCCM metric tool always has high cohesion and low coupling. The result states that MCCM metric always returns higher values comparing to LCC because, all the connection type between elements in the classes are taken under consideration. Moreover, Cohesion refers to the degree of the relatedness of the members in a component. High cohesion is a desirable property of software components. It is widely recognized that highly cohesive components tend to have high maintainability and reusability.

E. Comparison between two projects

This section presents measuring and comparison the quality of two projects by using MCCM tool.

- First project (car_sales_system.jar)

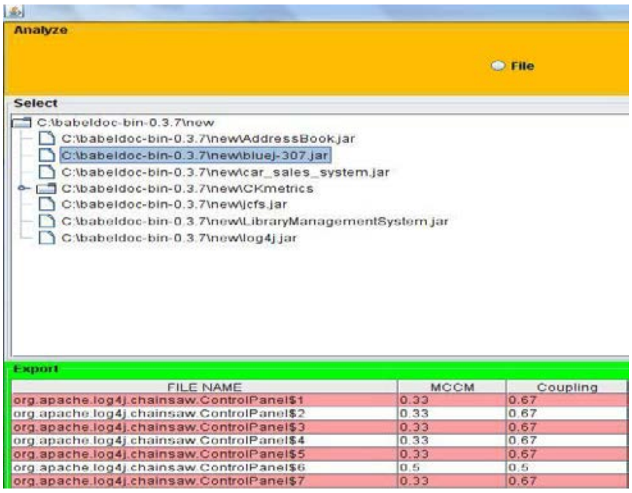


```
C:\babeldoc-bin-0.3.7\new
  C:\babeldoc-bin-0.3.7\new\AddressBook.jar
  C:\babeldoc-bin-0.3.7\new\bluej-307.jar
  C:\babeldoc-bin-0.3.7\new\car_sales_system.jar
  C:\babeldoc-bin-0.3.7\new\CKmetrics
  C:\babeldoc-bin-0.3.7\new\jcfs.jar
  C:\babeldoc-bin-0.3.7\new\LibraryManagementSystem.jar
  C:\babeldoc-bin-0.3.7\new\log4j.jar
```

FILE NAME	MCCM	Coupling
AboutDialog	0.88	0.12
AddCarPanel	0.83	0.17
Car	0.88	0.12
CarDetailsComponents	0.85	0.15
CarSalesSystem	0.95	0.05
CarsCollection	0.68	0.32
Manufacturer	0.58	0.42
SearchByAgePanel	0.72	0.28
SearchByOtherPanel	0.77	0.23
ShowAllCarsPanel	0.62	0.38
WelcomePanel	0.69	0.31

Fig.7.Results of car_sales_system.jar

- Second project (Bluej-307.jar)



```
C:\babeldoc-bin-0.3.7\new
  C:\babeldoc-bin-0.3.7\new\AddressBook.jar
  C:\babeldoc-bin-0.3.7\new\bluej-307.jar
  C:\babeldoc-bin-0.3.7\new\car_sales_system.jar
  C:\babeldoc-bin-0.3.7\new\CKmetrics
  C:\babeldoc-bin-0.3.7\new\jcfs.jar
  C:\babeldoc-bin-0.3.7\new\LibraryManagementSystem.jar
  C:\babeldoc-bin-0.3.7\new\log4j.jar
```

FILE NAME	MCCM	Coupling
org.apache.log4j.chainsaw.ControlPanel$1	0.33	0.67
org.apache.log4j.chainsaw.ControlPanel$2	0.33	0.67
org.apache.log4j.chainsaw.ControlPanel$3	0.33	0.67
org.apache.log4j.chainsaw.ControlPanel$4	0.33	0.67
org.apache.log4j.chainsaw.ControlPanel$5	0.33	0.67
org.apache.log4j.chainsaw.ControlPanel$6	0.5	0.5
org.apache.log4j.chainsaw.ControlPanel$7	0.33	0.67

Fig.8.Results of Bluej-307.jar

The displayed results in the above figures explaining, the first project (car_sales_system.jar) has higher value in (MCCM column) more than the second project (Bluej-307.jar), and has value in (Coupling column) is less than the second project (Bluej-307.jar). So we conclude that the first project (car_sales_system.jar) has higher quality more than the second project (Bluej-307.jar), because all classes in the first project (car_sales_system.jar) have high cohesion and low coupling, and all classes in the second project (Bluej-307.jar) have low cohesion and high coupling.

These mechanisms are presented in Table (5) along with the cohesion metrics [5].

From Table (5) the following conclusions are drawn:

- There are two ways via which`h the interactions among methods can be captured :

(1) Method Invocation (2) Attribute sharing.

- The most effective Interaction Type is M → M. This may be considered as the best way to capture the cohesion of a class because methods play a better role (than attributes) in determining what the functionality of a class.
- It is interesting to note that with the mechanisms presented in Table (5), the cohesion metrics correlations can easily be explained. For instance: MCCM use all the stronger interaction type and interaction mode, whereas LCC does not.

V. Conclusion

We have developed a cohesion measurement tool for Java software to automate the computation of the major existing class cohesion metrics including ours. In order to demonstrate the effectiveness of the MCCM cohesion metric, we performed a case study on several systems.

The obtained results confirm our hypothesis. They show clearly that the MCCM metrics, based on a combinationof the proposed criteria, capture more pairs of connected methods than the existing cohesion metrics, particularly the ones supposed implicitlytaking into account the interactions between methods (Method Invocation and Attribute Usage). We believe that the present work constitutes an improvement of class cohesion assessment.

References

[1] Tejdeda Alhussen Alhadi, Dr Omer Jomah, Xavier Patrick kishore, Sagaya Aurelia, Mathematical Model of Object Oriented Class Cohesion Metric MCCM, unpublished

[2]Lars Vogel, “Eclipse JFace Overview”, Version 2.7, October 2012, From http://www.eclips.org/platform,.

[3] H. S. Chae, Y. R. Kwon and D H. Bae, A cohesion measure for object-oriented classes, Software Practice and Experience, No. 30, pp. 1405-1431, 2000.

[4] Al Dallal, J. and Morasca, S., Predicting Object-Oriented Class Reusability Using Internal Quality Attributes, Empirical Software Engineering, in press, 2012.

[5] Abubakar A., “Implementation and validation of object-oriented design –Level cohesion metrics”, Thesis presented, Dhahran, Saudi Arabia, Computer Science, January 2005.

Ms. Tejdeda Alhussen Alhadi (Feburary 1, 1980) is now with Faculty of Education, Azzaytuna University, Bani-walid, Libya. She is into teaching profession for more than 13 years. She has done B.Sc and M.Sc in Computer Science from Libyan Academy. She has also been involved in various administration related activities. Her specialization includes Database, Software Engineering and Artifical intelligence. She has published various national and international papers and guided many projects.

Dr. Omer Saleh Mahmod Jamah (January 25,1973) is now the Director of Post graduate cum Research and Development and Head of the department of Computer science, Faculty of education, Azzaytuna university, Baniwalid, Libya. He received his B.Sc. in Control

System and Measurement (1995), M.Sc. in Electrical and Computer Measurement (2004), and Ph.D. in Electrical engineering, Automatics computer science and electronics from AGH University of technology, Krakow, Poland. He has done his Diploma in Planning and time management from Canada Global Centre, Canada. Now he is heading Computer Science department, Faculty of Education, Azzaytuna University, Baniwalid, Libya. His research interest includes multicriteria optimization for solving optimal control problems and Fuzzy logic. He has published 12 papers and attended various national and international Level conferences and workshops.

Mr. Xavier Patrick Kishore (November 6, 1973) received his BSc Mathematics (1994), Master of Computer Application (2002) and Diplomas in E-Commerce and Advanced software Technology. He has received Brain bench certification in Java and HTML. Now he is working in Department of computer science Faculty of Education, Azzaytuna University,Baniwalid, Libya. He is specializedin programming languages. His current research interest includes Natural languageprocessing. He has authored more than 9 papers and attended many conferences.

Er. Mrs. Sagaya Aurelia(November 9,1978) par-time research scholar in Bharathidasan university . Now she is with department of Computer Science, Faculty of Education, Azzaytuna University, Bani-walid, Libya. She received her Diploma in Electronics and Communication (1997),B.E (Bachelor of Engineering specialized in Electronics and Communication Engineering(2000) and M.Tech in Information Technology(2004),she has also done her Postgraduation diplomasinBusiness Administration (PGDBA) and Journalism and Mass Communication(PGDJMC). She has received Brainbench certification in HTML. Her current research interest includes Virtual reality, Augmented reality and Human Computer Interaction and User interface Design. She has authored14 papers and attendance several national and international level workshops and conferences.

Can One-Chip Parallel Computing *Be Liberated From* Ad Hoc Solutions? *A* Computation Model Based Approach *and Its* Implementation

Gheorghe M. Ştefan and Mihaela Maliţa

In July 2010 David Patterson said in *IEEE Spectrum* that *"the semiconductor industry threw the equivalent of a Hail Mary pass when it switched from making microprocessors run faster to putting more of them on a chip – doing so without any clear notion of how such devices would in general be programmed"* warning us that one-chip parallel computing seems to be in trouble. Faced with the problems generated by all those *ad hoc* solutions, we propose a fresh restart of parallel computation based on the synergetic interaction between: (1) a parallel computing model (Kleene's model of partial recursive functions), (2) an abstract machine model, (3) an adequate architecture and a friendly programming environment (based on Backus's FP Systems) and (4) a simple and efficient generic structure. This structure is featured with an *Integral Parallel Architecture*, able to perform all the five forms of parallelism (data-, reduction-, speculative-, time- and thread-parallelism) which result from Stephen Kleene's model and is programmed as John Backus dreamed. Our first embodiment of a one-chip parallel generic structure is centered on the cellular engine *ConnexArray*^{TM} which is part of the SoC BA1024 designed for HDTV applications. On real chips we measured $6\,GOPS/mm^2$ and $120\,GOPS/Watt$ peak performance.

***Index Terms*—Parallel computing, recursive functions, parallel architecture, functional programming, integral parallel computation.**

I. Introduction

It seems that the emergence of parallelism brings difficult times for computer users who lack a friendly environment to develop their applications. But the situation is not new. In 1978 John Backus complained in a similar manner, in [4], telling us that *"programming languages appear to be in trouble"*. In the following fragment he coined the term *"von Neumann bottleneck"*, trying to explain the main limitation of the sequential programming model, dominant in his time:

> *"Von Neumann programming languages use variables to imitate the computer's storage cells; control statements elaborate its jump and test instructions; and assignment statements imitate its fetching, storing, and arithmetic. The assignment statement is the von Neumann bottleneck of programming languages and keeps us thinking in word-at-a-time terms in much the same way the computer's bottleneck does."*

In his seminal paper John Backus proposes two important things: (1) the main limitation of sequential computing, proposing its *PF Systems*, a new programming style, and (2) a formal definition of a generic parallel architecture. Removing the "von Neumann bottleneck" means not only freeing the programming style from parasitic control actions, but also opens the way for triggering parallel actions on large amount of data avoiding explicit cumbersome data and code manipulations.

The history of parallel computing ignored the second suggestion offered by Backus. The parallel computation already begun wrong, with *ad hoc*, speculative constructs, considering that more than one machine, more or less sophisticatedly interconnected, will have the force to solve the continuously increasing hunger for computing power. The scientific community was from the beginning too much focused on parallel hardware and parallel software, instead of solving first the computational model and architectural issues. In fact our computing community started from building too early parallel hardware and then learned quickly that we are not able to program it efficiently. There are few errors in this approach.

First of all, putting together 4, 8 or 128 processors does not mean necessarily that we built a parallel machine. A parallel machine must be thought as a n-cell system, where n is a however large number. Scalability must be the main feature of a parallel engine.

Second, a number of n Turing-based sequential machines, interconnected in a certain network can not offer a starting point in designing a parallel computer, because the role of the interconnections could be more important and complex than the effects of the cells they interconnect.

Third, while the mono-processor computer is theoretically grounded in a *computing model*[1], is based on an *abstract machine model*[2] and is supported by an appropriate *architectural approach*[3], the multi- or many-processor approach is not yet based on an appropriate computational model, there is no a validated abstract machine model or a stabilized architectural environment which refers to a n-sized computational mechanism. Indeed, the sequential, mono-processor computer was backed, by turn, by Turing's (or equivalent) *computing*

Gheorghe M. Ştefan is with the Department of Electronic Devices, Circuits and Architectures, Politehnica University of Bucharest, Bucharest, Romania, e-mail: gstefan@arh.pub.ro

Mihaela Maliţa is with Department of Computer Science, Saint Anselm College, Manchester, NH, e-mail: mmalita@anselm.edu

[1] A ***computing model*** is a *mathematical definition* for automatic computing provided by the theory of computation originated during the 1930s from the seminal ideas triggered by Kurt Gödel in the works of Alonzo Church, Stephen Kleene, Emil Post and Alan Turing.

[2] An ***abstract machine model*** is a *structural definition* which provides the computer organization able to implement the computation defined by a computing model.

[3] An ***architecture*** provides a clear segregation between the hardware and software, using the interface offered by the functionality of an appropriate instruction set.

ISBN: 978-1-61804-237-8

model, by the von Neumann or Harvard *abstract machine model*, and later by an *appropriate architectural approach* when the complexity of hardware-software interaction became embarrassing.

We believe that considering a parallel machine as an *ad hoc* collection of already known sequential machines "appropriately" interconnected is the worst way to start thinking about parallel computation. Maybe, parallel computing is the natural way to make computation and the sequential computation is an early stage of computation we were obliged to accept because of obvious technological limitations, and now is the time to restart the process in the current, improved technological conditions.

A. What is Wrong with Parallel Computing?

There is a big difference between the history of how the sequential computing domain emerged and what happened with the parallel computing domain in the last half century. In the first case there is a coherent sequence of events leading to the current stage of sequential computation, while for parallel computation the history looks like a chaotic flow of events. Let us schematize the emergence of the two sub-domains of computing. First, for sequential computation we have:

- **1936 – computational models** : four equivalent models are published [33] [7] [14] [23] (all reprinted in [9]), out of which the *Turing Machine* offered the most expressive and technologically appropriate suggestion for future developments
- **1944-45 – abstract machine models** : MARK 1 computer, built by IBM for Harvard University, consecrated the term *Harvard abstract model*, while von Neumann's report [34] introduced what we call now the *von Neumann abstract model*; these two concepts backed the *RAM* (random access machine) abstract model used to evaluate algorithms for sequential machines
- **1953 – manufacturing in quantity** : IBM launched *IBM 701*, the first large-scale electronic computer
- **1964 – computer architecture** : in [6] the concept of *computer architecture* (low level machine model) is introduced to allow independent evolution for the two different aspects of computer design, which have different rate of evolution: software and hardware; thus, there are now on the market few stable and successful architectures, such as x86, ARM, PowerPC.

Thus, in a quarter of century, from 1936 to the early 1960s, the sequential computer domain evolved coherently from theoretical models to mature market products.

Let's see now what happened in the parallel computing domain:

- **1962 – manufacturing in quantity** : the first symmetrical MIMD engine is introduced on the computer market by Burroughs
- **1965 – architectural issues** : Edsger W. Dijkstra formulates in [10] the first concerns about specific parallel programming issues
- **1974-76 – abstract machine models** : proposals of the first abstract models (bit vector models in [24], [25], and PRAM models in [11], [12]) start to come in after almost two decades of non-systematic experiments (started in the late 1950) and too early market production
- **? – computation model** : no one yet considered it, although it is there waiting for us (it is about Kleene's model [14]).

Now, in the second decade of the 3rd millennium, after more than half century of chaotic development, it is obvious that ***the history of parallel computing is distorted by missing stages and uncorrelated evolutions***[4]. The domain of what we call parallel computation is unable to provide a stable, efficient and friendly environment for a sustainable market.

In the history of parallel computation the stage of defining the parallel computational model is skipped, the stage of defining the abstract machine model is too much delayed and confused with the definition of the computation model, while we do not have yet a stable solution for a parallel architecture. Because of this incoherent evolution even the parallel abstract models, by far the most elaborated topics in parallel computation, are characterized by a high degree of artificiality, due to their too speculative character.

B. Parallel Abstract Machine Models

What we call today parallel computation models are in fact a sort of abstract machine models, because true computational models are about how computable functions are defined, not (unrealistic) hardware constructs which interconnect sequential computing engines. Let us take a look in the world of parallel abstract machine models.

1) Parallel Random Access Machine – PRAM

The PRAM abstract model is considered, in [13], a "natural generalization" of the Random Access Machine (RAM) abstract model. It is proposed in [11] and in [12], and consists of n processors and a m-module shared memory, with both, n and m of unbounded size. Each processor has its own local memory. A memory access is executed in one unit time, and all the operations are executed in one unit time. The access type a processor has to the shared memory differentiates four types of PRAMs: EREW (exclusive read, exclusive write), CREW (concurrent read, exclusive write), ERCW (exclusive read, concurrent write), CRCW (concurrent read, concurrent write). The flavor of this taxonomy is too structural, somehow speculative and artificial, unrelated directly with the idea of computation, besides it refers to unrealistic mechanisms. The model is a collection of machines, memories and switches, instead of a functionally oriented mechanism as we have in the Turing Machine, lambda-calculus, recursive functions models. More, as an abstract machine model it is unable to provide:

- accurate predictions about the effective performances related to time, space, energy, because of the unrealistic structural suppositions it takes into account
- a lead to programming languages, because of the fragmented cellular approach which associates a programming language only at the cell level, without any projection to the system level

[4] In this paper the economical, social, psychological aspects are completely ignored, not because they are irrelevant, but because we concentrate only on the pure technical aspects of the problem.

- any realistic embodiment suggestion, because it ignores any attempt to provide details about memory hierarchy, interconnection network, communication, ...

Ignoring or treating superficially details about communication and memory hierarchy is deceptive. For example: pure theoretic model for RAM says $n \times n$ matrix multiplication time is in $O(n^3)$, but real experiments (see [26]) provide $O(n^{4.7})$. If for such a simple model and algorithm the effect is so big, then we can imagine the disaster for the PRAM model on really complex problems!

2) Parallel Memory Hierarchy – PHM

The PHM model is also a "generalization", but this time of the Memory Hierarchy model applied to the RAM model. This version of the PRAM model is published in [2]. The computing system is hierarchically organized on few levels and in each node the computation is broken in many independent tasks distributed to the children nodes.

3) Bulk Synchronous Parallel – BSP

The BSP model divides the program in *super-steps* [35]. Each processor executes a *super-step*, which consists of a number of computational steps using data stored in their own local memories. At the end of the super-step processors synchronize data by message passing mechanisms.

4) Latency-overhead-gap-Processors – LogP

The *LogP* model is designed to model the communication cost in a parallel engine [8]. The parameters used to name and to define the model are: *latency* – L – time for a message to move from a processor to another; *overhead* – o – time any processor spends for sending or receiving a message; *gap* – g – is the minimum time between messages; the number of *processors* – P –, each having a big local memory. The first three parameters are measured in clock cycles. The model is able to provide an evaluation which takes into account the communication costs in the system.

The last three models are improved forms of the PRAM model; they provide a more accurate image about parallel computation, but all of them inherit the main limitation of the mother model, the too speculative and artificial PRAM model. The general opinions about PRAM are not very favorable:

> *"Although the PRAM model is a natural parallel extension of the RAM model, it is not obvious that the model is actually reasonable. That is, does the PRAM model correspond, in capability and cost, to a physically implementable device? Is it fair to allow unbounded numbers of processors and memory cells? How reasonable is it to have unbounded size integers in memory cells? Is it sufficient to simply have a unit charge for the basic operations? Is it possible to have unbounded numbers of processors accessing any portion of shared memory for only unit cost? Is synchronous execution of one instruction on each processor in unit time realistic?"* ([13], p. 26)

Maybe even the authors of the previous quotation are somehow wrong, because parallel computation modelled by PRAM is not a natural extension of the sequential computation, on the contrary, we believe that the sequential computation is a special case of parallel computation.

We are obliged to assert that the PRAM model, and its various versions[5], have little, if any, practical significance. In addition, the delayed occurrence of these models, after real improvised parallel machines were already on the market, have a negative impact on the development of the parallel computation domain.

C. What Must be Done?

The question *What must be done?* is answered by *We must restart as we successfully started for sequential computing*. And we have also good news: a lot of stuff we need is there waiting to be used. There is a ***computational model*** which is a perfect match for the first step in the emergence of the parallel computation: the *partial recursive functions* model proposed by Stephen Kleene in 1936, the same year in which Turing, Church and Post made their proposals. It can be used, in a second step, to derive from it an ***abstract machine model*** for parallel computation. For the third step, the *FP Systems*, proposed in 1978 by John Backus, are waiting to be used in order to provide the ***architecture*** or the ***low level model*** for parallel computation.

However, we learned a lot from the work already done for developing abstract models for parallelism. The evaluation made in [17] provides the main characteristics to be considered in the definition of any abstract machine model:

- ***Computational Parallelism*** must be performed using the *simplest* and *smallest* cells; in order to increase the area and energy efficiency, the structural granularity must decrease with the number of cells (see [27])
- ***Execution Synchronization*** can be maintained simple only if the computational granularity is small, preferably minimal
- ***Network Topology*** is suggested by the parallel computational model and must be kept as simple as possible; communication depends on many other aspects and the optimization is a long and complex process which will provide only in time insights about how the network topology must be structured and tuned
- ***Memory Hierarchy*** is a must, but the initial version of the abstract model is preferably to have only a minimal hierarchy; it can be detailed only as a consequence of an intense use in various application domains; the question *caches or buffers?* has not yet an unanimously accepted answer
- ***Communication Bandwidth*** being very costly, in size and energy, must be carefully optimized taking into account all the other six characteristics
- ***Communication Latency*** can be hidden by carefully designed algorithms and by disconnecting, as much as possible, the computation processes by the communication processes

[5] Besides the discussed ones there are many others, like: Asynchronous PRAM, XPRAM (performs periodic synchronization), LPRAM (includes memory latency costs), BPRAM (block transfer oriented), DRAM (adds the level of distributed local memory), PRAM(m) (limits the size of the globally shared memory to m).

- ***Communication Overhead*** has a small impact only for simple communication mechanisms involving simple network, simple synchronization, simple cells, small hierarchy.

Obviously, the complexity can not be avoided eventually, but the way complexity manifests can not be predicted, it must be gradually discovered after a long time use of a simple generic model.

Unfortunately, the emergence of parallel computing occurred in a too dynamic and hurried world, with no time to follow the right path. A new restart is required in order to define a simple generic parallel machine, *subject to organic improvements in a long term evaluation process*. Our proposal is a five-stage approach:

1) use Kleene's partial recursive functions as the ***parallel computational model*** to provide the theoretical framework
2) define the ***abstract machine model*** using meaningful forms derived from Kleene's model
3) put on top of the abstract machine model a ***low level (architectural) model*** description based on Backus's FP Systems
4) provide the simplest ***generic parallel structure*** able to run the functions requested by the low level model
5) ***evaluate*** the options made in the previous three steps in the context of *the computational motifs* highlighted by *Berkeley's View* in [3], looking mainly for systematic or local weaknesses of the architectural model or generic structure in implementing typical algorithms.

The first two steps will be completed in the next two sections, followed by a section used to sketch only the third stage. For the fourth stage an already implemented engine is presented. The fifth step will be only shortly reviewed; it is left for future work. Only after the completion of this 5-stage project the discussion on parallel programming models can be started based on a solid foundation.

II. Kleene's Model is a Parallel Computational Model

Kleene's model of partial recursive functions contains the parallel aspects of computation in its first rule – the ***composition rule*** –, while the next two rules – *primitive recursion* and *minimalization* – are elaborated forms of composition (see [18]). Thus, composition captures directly the process of parallel computing in a number of $p+1$ functions by:

$$f(x_1,\ldots,x_n) = g(h_1(x_1,\ldots,x_n),\ldots,h_p(x_1,\ldots,x_n))$$

where, there are involved, at the first level, p functions – $h_1,\ldots,h_p$ – and a p-variable reduction function g (sometimes implementable as a $(log\, p)$-depth binary tree of $p-1$ functions).

In Figure 1 the ***circuit embodiment*** of the composition rule is represented. It consists of a layer of p cells (each performs the function h_i, for $i=1,2,\ldots,p$) and a module which performs the reduction function g. In the general case the values of the input variables are sent to all the p cells performing functions h_i and the resulting p-component vector $\{h_1(x_1,\ldots,x_n),\ldots,h_p(x_1,\ldots,x_n)\}$ is reduced to a scalar by the module g.

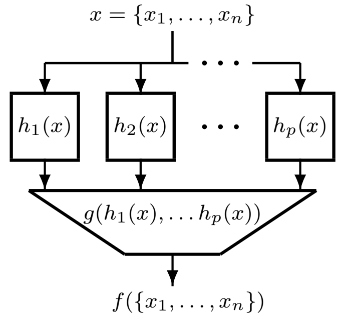


Fig. 1. **The circuit structure associated to composition.**

Two kinds of parallelism are foreseen at this level of our approach: a *synchronic* parallelism – on the first layer of cells – and a *diachronic* (pipeline) parallelism – between the two levels of the structure.

The partial recursive functions model uses two other rules: the *primitive recursive rule* and the *minimalization rule*. We will prove that both can be defined composing special forms of the composition rule. Thus, we will conclude that the composition rule could be the only mechanism to be considered in describing the parallel computation.

A. Reducing Primitive Recursion to Composition

In this subsection is proved that the second rule of the partial recursive model of computation, the primitive recursive rule, is reducible to the repeated application of specific forms of composition rule. Let be the composition:

$$C_i(x_1,\ldots,x_i) = g(f_1(x_1,\ldots,x_i),\ldots,f_{i+1}(x_1,\ldots,x_i))$$

If g is the identity function $g(y_1,\ldots,y_{i+1}) = \{y_1,\ldots,y_{i+1}\}$ and $f_1(x_1,\ldots,x_i) = h_i(x_1)$, $f_2(x_1,\ldots,x_i) = x_1$, …, $f_{i+1}(x_1,\ldots,x_i) = x_i$, then

$$C_i(x_1,\ldots,x_i) = \{h_i(x_1),x_1,x_2,\ldots,x_i\}$$

The repeated application of C_i (see Figure 2a), starting from $i=1$ with $x_1 = x$ allows us to compute the pipelined function P (see Figure 2b):
$P(x) = \{h_1(x),h_2(h_1(x)),h_3(h_2(h_1(x))),\ldots$
$\ldots,h_k(h_{k-1}(\ldots(h_1(x)\ldots)),\ldots\}$

The function $P(x)$ is a total function if the functions h_i are total functions and it is computed using only the repeated application of the composition rule.

The primitive recursion rule defines the function $f(x,y)$ using the expression

$$f(x,y) = g(x,f(x,(y-1)))$$

where $f(x,0)=h(x)$. The iterative evaluation of the function f is done using the following expression:

$$f(x,y) = \underbrace{g(x,g(x,g(x,\ldots g}_{y\,times}(x,h(x))\ldots)))$$

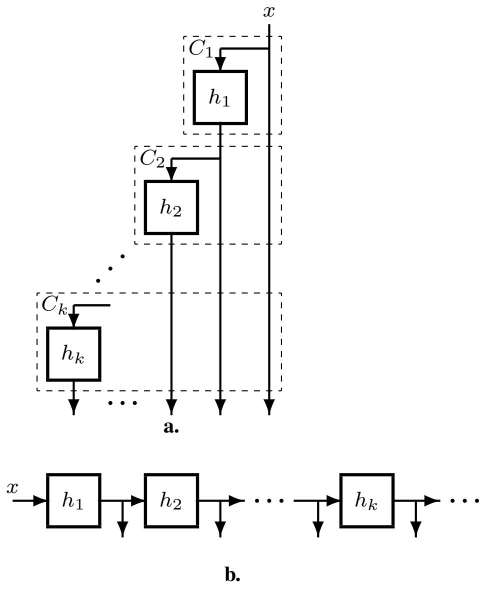


Fig. 2. **The pipeline structure as a repeated application of the composition C_i. a. The explicit application of C_i. b. The resulting multi-output pipelined circuit structure P.**

In Figure 3 is represented the iterative version of the structure associated to the primitive recursive rule. The functions used in the iterative evaluation are:

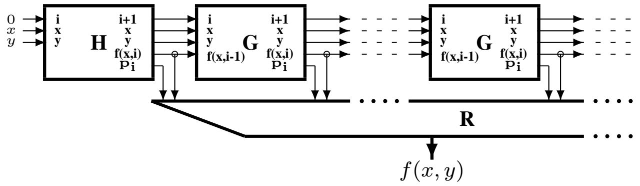


Fig. 3. **The circuit which performs the partial recursive computation.**

- $H(i,x,y) = \{(i{+}1), x, y, f(x,0), p_i\}$, receives the index $i = 0$ and the two input variables, x and y, and returns: the incremented index, $i+1$, the two input variables, $f(x,i)$, which is $h(x)$, and the predicate $p_i = p_0 = (y == 0)$. The predicate and the value of the function are used by the reduction function R, while the next function in pipe, G_1, receives $\{(i+1), x, y, f(x,0)\}$.
- $G(i,x,y,f(x,(i-1))) = \{(i+1), x, y, f(x,i), p_i\}$ receives the index i, the two input variables, x and y, $f(x,(i-1))$, and returns: the incremented index, $i+1$, the two input variables, $f(x,i)$, and the predicate $p_i = (y == i)$.
- $R(\{\{p_0, f(x,0)\}, \{p_1, f(x,1)\}, \ldots, \{p_i, f(x,i)\}, \ldots\}) =$ $IP(trans(\{\{p_0, f(x,0)\}, \{p_1, f(x,1)\}, \ldots$ $\ldots, \{p_i, f(x,i)\}, \ldots\})) =$ $IP(\{p_0, p_1, \ldots, p_i, \ldots\}, \{f(x,0), f(x,1), \ldots$ $\ldots, f(x,i), \ldots\}) = f(x,y)$ is a reduction function; it receives a vector of pairs predicate-value, of form $\{(y == i), f(x,i)\}$, and returns the value whose predicate is `true`. Function R is a composition of two functions: $trans$ (transpose), and IP (inner product). Both are simple functions computed by composition.

The two stage computation just described, as a structure indefinitely extensible to the right, is a theoretical model, because the index i takes values no matter how large, similar with the indefinitely extensible ("infinite") tape of Turing's machine. But, it is very important that the algorithmic complexity of the description is in $O(1)$, because the functions H, G and R have constant size descriptions.

B. Reducing Minimalization to Composition

In this subsection is proved that the third rule of the partial recursive model of computation, the minimalization rule, is also reducible to the repeated application of specific forms of the composition rule.

The minimalization rule computes the value of $f(x)$ as the smallest y for which $g(x,y) = 0$. The algorithmic steps used in the evaluation of function $f(x)$ consist of 4 reduction-less compositions and a final reduction composition, as follows:

1) $f_1(x) = \{h_0^1(x), \ldots h_i^1(x), \ldots\} = X_1$, with $h_i^1(x) = \{x, i\}$
2) $f_2(X_1) = \{h_0^2(X_1), \ldots h_i^2(X_1), \ldots\} = X_2$, with $h_i^2(X_1) = \{i, p_i\}$, where $p_i = (g(sel(i, X_1)) == 0)$ is the predicate indicating if $g(x,i) = 0$, and sel is the basic function *selection* in Kleene's definition; provides pairs index-predicate having the predicate equal with 1 where the function g takes the value 0
3) $f_3(X_2) = \{h_0^3(X_2), \ldots h_i^3(X_2), \ldots\} = X_3$, with $h_i^3(X_2) = \{i, pref_i\}$, where $\{pref_0, \ldots pref_i, \ldots\} = prefixOR(p_0, \ldots p_i, \ldots)$; in [15] is provided a $O(log\ n)$ steps optimal recursive algorithm for computing the prefix function for n inputs
4) $f_4(X_3) = \{h_0^4(X_3), \ldots h_i^4(X_3), \ldots\} = X_4$, with $h_i^4(X_3) = \{i, ADN(pref_i, NOT(pref_{i-1}))\} = \{i, first_i\}$; provides pairs index-predicate where only the first occurrence, if any, of $\{i, 1\}$ is maintained, all the others take the form $\{i, 0\}$
5) $f_5(X_4) = R(\{\{first_0, 0\}, \ldots, \{first_i, i\}, \ldots\}) =$ $\{OR(\{first_0, \ldots, first_i, \ldots\}), IP(\{first_0, \ldots$ $\ldots, first_i, \ldots\}, \{0, \ldots, i, \ldots\})\} =$ $\{p, f(x)\} = p\ ?\ f(x)\ :\ -$ is a reduction function; it receives a vector of pairs predicate-value, of form $\{(y == i), f(x,i)\}$, and returns the value whose predicate is `true`, **if any**. If $p = 0$, then the function has no value.

The computation just described is also a theoretical model, because the index i has an indefinitely large value. But, the size of algorithmic description remains $O(1)$, because the functions f_j are completely defined by the associated generic functions h_i^j, for $j = 1, 2, 3, 4$.

C. Concluding about Kleene's Model

In this section we proved that the model of partial recursive functions can be expressed using only the composition rule, because the other two rules – primitive recursion and minimalization – are reducible to multiple applications of specific compositions. The resulting computational model is an ***intrinsic parallel model of computation***. The only rule defining it – the composition rule – provides two kinds of parallelism: the *synchronic parallelism* on the first stage of $h_i(x)$ functions, and a *diachronic parallelism* between the first stage and the reduction stage. (The reduction stage can be expressed in turn using log-stage applications of the composition rule.)

Thus, Kleene's model of parallel computation is described by the circuit construct represented in Figure 1, where p has value no matter how large. For a theoretical model it does not hurt. The *abstract model of parallel computation*, introduced in the next section, aims to remove the theoretical freedom allowed by the "infinite" physical resources.

The theoretical ***degree of parallelism***, δ, emphasized for the two-level function

$$f(x_1,\ldots,x_n) = g(h_1(x_1,\ldots,x_n),\ldots,h_p(x_1,\ldots,x_n))$$

is p for the first level of computation, if $h_i(x_1,\ldots,x_n)$ are considered atomic functions for $i = 1,\ldots p$, while for the second level δ is given by the actual description of the p-variable function g. The theoretical degree of parallelism depends on the possibility to provide the most detailed description as a composition using atomic functions.

Informally, we ***conjecture*** that *the degree of parallelism for a given function f, δ_f, is the sum of the degree o parallelism found on each level divided by the number of levels*. Therefore, theoretically the function f can be computed in parallel only if $\delta_f > 1$.

For example, if $f(x_1,\ldots,x_n) = \sum_{i=1}^{n} x_i^2$ is the inner product of a vector with itself, then the first level of computation is *data-parallel* with $h_i = x_i^2$ and the second level of computation, the function g, is a *reduction-parallel* function computed by a log-depth binary tree of two-input adders. If *multiply* and *add* are considered atomic operations and n a power of 2, then the value of δ for f is:

$$\delta_f = (n + n/2 + n/4 + \ldots + 1)/(1 + log_2 n) =$$
$$(2n-1)/(1 + log_2 n) \in O(n/log\, n)$$

It seems that a degree of parallelism $\delta \in O(n/log\, n)$ is the lower limit for what we can call a *reasonable efficient parallel computation*.

The composition rule will be considered as the starting point, in the next section, for defining an abstract parallel machine model.

III. An Abstract Parallel Machine Model

The distance from Turing's model to Harvard or von Neumann models is the distance between a ***mathematical*** *computational model* and an *abstract* ***machine*** *model*. The first model is mainly about ***What*** *is computation?* and the second is more about ***How*** *computation is done?*. Our abstract machine model takes into consideration some meaningful simplified forms of the composition rule. We claim that the following five forms of composition provide the structural requirements for a *reasonable* efficient parallel engine.

A. Meaningful Simplified Forms of Compositions

1) Data-parallel

The first simplified composition distributes along the functionally identical cells the input sequence of data $\{x_1,\ldots,x_p\}$, and considers that the second level executes the identity function, i.e., $h_i(x_1,\ldots,x_p) = h(x_i)$ and $g(y_1,\ldots,y_p) = \{y_1,\ldots,y_p\}$. Then,

$$f(x_1,\ldots,x_p) = \{h(x_1),\ldots,h(x_p)\}$$

where $x_i = \{x_{i1},\ldots,x_{im}\}$ are sequences of data, is a ***data-parallel*** computation.

A more complex data-parallel operation is the conditioned (predicated) execution:
$f(\{x_1,\ldots,x_p\},\{b_1,\ldots,b_p\}) =$
$\{(b_1\ ?\ h_T(x_1) : h_F(x_1)),\ldots,(b_p\ ?\ h_T(x_p) : h_F(x_p))\}$
where: b_i are Boolean variables.

The circuit associated with the data-parallel computation is a cellular structure (see Figure 4a), where each cell receives its own component x_i from the input sequence. The execution is *unconditioned* – each cell executes: $h(x_i)$ –, or it is *conditioned* by the state of the cell, expressed by locally computed Booleans, and each cell executes: $b_i\ ?\ h_T(x_i) : h_F(x_i)$. Each cell must have local memory for storing the sequence x_i, for the working space and data buffers. The sequence of operations performed by the array of cells is stored in the program memory of the associated control circuit.

2) Reduction-parallel

While the first type of parallelism assumes that the reduction function is the identity function, the second form makes the opposite assumption: the first layer, of the synchronous parallelism, contains the identity functions: $h_i(x_i) = x_i$. Thus the general form becomes:

$$f(x_1,\ldots,x_p) = g(x_1,\ldots,x_p)$$

which *reduces* the input sequence of variables to a single variable (see Figure 4b). The circuit organization of the reduction is tree-like. It consists of a repeated application of various compositions. The size of the associated structure is in the same range as for the data-parallel, while the depth is in $O(log\ p)$.

Because, in the current applications there are only few meaningful reduction functions, the reduction-parallel operations are usually performed using circuits instead of programmable structures.

3) Speculative-parallel

The third simplified composition is somehow complementary to the first: the functionally different cells – h_i – receive the same input variable – x – while the reduction section is the same. Then,

$$f(x) = \{h_1(x),\ldots,h_p(x)\}$$

where: x is a sequence of data. The function returns a sequence of sequences.

There are two ways to differentiate the functions $h_i(x)$:

1) $h_i(x)$: represents a specific sequence of operation for each i. Then, the local memory in each cell contains, besides data, the sequence of operations

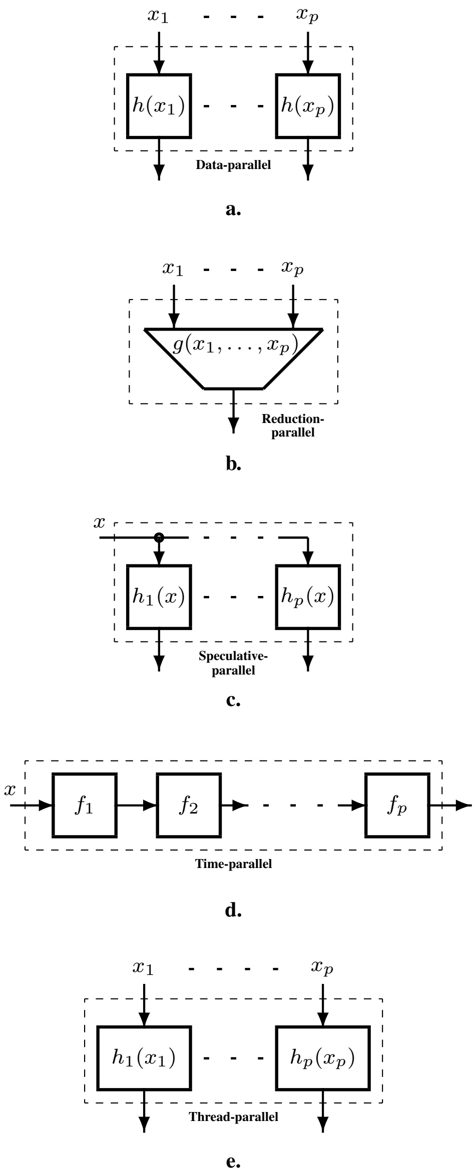


Fig. 4. **Five types of parallelism as particular forms of composition (see Figure 1)**

2) $h_i(x) = g(i, x)$: the sequence of operations are identical for each i, but the function has the parameter i. Then, the local memory contains only data, and the execution takes into account the index of the cell to differentiate the local execution of the sequence of operations stored in the memory of the associated control device.

The circuit associated to the speculative-parallel computation is a cellular structure (see Figure 4c), each cell receiving the same input variable – x – which is used to compute different functions. The general case of speculative-parallel computation requests local data and program memory. While the data-parallel cell is an *execution unit*, the speculative-parallel cell is sometimes a *processing unit*.

4) Time-parallel

There is the special case when the functions are defined for $p = 1$, i.e., $f(x) = g(h(x))$. Then, here is no synchronous parallelism. Only time (diachronic), pipelined parallelism is possible if in each "cycle" a new value is applied to the input. Thus, in each "cycle" the function h is applied to $x(t)$ (which is x at the "moment" t) and g is applied to $h(x(t-1))$ (where $x(t-1)$ is the value applied to the input at the "moment" $t-1$). The system delivers in each "cycle" the result of a computation supposed to be performed in 2 "cycles", or we say that the system works in parallel for computing the function f for 2 successive values.

Many applications of $f(x) = g(h(x))$ result in the m-level *"pipe"* of functions:

$$f(x) = f_m(f_{m-1}(\ldots f_1(x)\ldots))$$

where: x is an element in a stream of data. The resulting structure (see Figure 4d) is a parallel one if in each "cycle" a new value for x is inserted in the pipe, i.e., it is applied to f_1.

This type of parallelism comes with a price: the *latency* time, expressed in number of cycles, between the insertion of the first value and the occurrence of the corresponding result.

5) Thread-parallel

The last simplified form of composition, we consider for our abstract machine model, is the most commonly used in current real products. It is in fact the simplest parallelism, applied when the solution of a function is a sequence of objects computed completely independent. If $h_i(x_1, \ldots, x_n) = h_i(x_i)$ and $g(h_1, \ldots, h_p) = \{h_1, \ldots, h_p\}$, then the general form of composition is reduced to:

$$f(x_1, \ldots, x_p) = \{h_1(x_1), \ldots, h_p(x_p)\}$$

where: x_i is an sequence of data. Each $h_i(x_i)$ represents a distinct and independent *thread* of computation performed in distinct and independent cells (see Figure 4e). Each cell has its own data and program memory.

B. Integral Parallelism

We make the assumption that the previously described particular forms of composition – the only rule that we showed is needed for the calculation of any partial recursive function – cover the features requested for a ***parallel abstract machine model***. This assumption remains to be (at least partially) validated in the fifth stage of our proposal, which evaluates the model against all the known computational motifs.

1) Complex vs. Intense in Parallel Computation

The five forms of parallelism previously emphasized perform two distinct types of computation:

- **intense computation** : the algorithmic complexity of the function $f(x_1, \ldots, x_n)$ is constant, while the size of data is in $O(F(n))$
- **complex computation** : the algorithmic complexity of the function is in $O(F(n))$.

Revisiting the types of parallel computation we find that:

- ***data-parallel*** computation is ***intense***, because it is defined by one function, h, on many data, $\{x_1, \ldots, x_p\}$

- ***reduction-parallel*** computation is ***intense*** because the function is simple (constant size definition) and the size of data is in $O(p)$
- ***speculative-parallel*** computation ***most of time is intense*** while ***sometimes is complex***, because:
  - when $h_i(x) = g(i,x)$ the resulting computation is intense, because the functional description has constant size, while the data is $\{x, 0, 1, \ldots, p\}$; we call it *intense speculative-parallel* computation
  - when $h_i \neq h_j$ for $i, j \in \{0, 1, \ldots, p\}$ the functional description has the size in $O(p)$ and the size of data is in $O(1)$; we call it *complex speculative-parallel* computation.
- ***time-parallel*** computation is ***most of the time complex***, because the definition of the m functions f_i has the size in $O(m)$, while ***sometimes is intense***, when the pipe of functions is defined using only a constant number of different functions
- ***thread-parallel*** computation is ***complex*** because the size of the description for $h_i \neq h_j$ for $i, j \in \{1, \ldots, p\}$ is in $O(p)$.

2) Many-Core vs. Multi-Core

For a general purpose computational engine all the five forms must be supported by an integrated abstract model. We call the resulting model: ***integral parallel*** *abstract machine model* (see Figure 5).

In Figure 5 there are emphasized two sections, called MANY-CORE and MULTI-CORE. The first section contains the cells $c_1, \ldots c_p$ and the log-depth network **redLoopNet** (which performs reduction functions and closes a global loop responsible for performing *scan* functions). They are used for data-, speculative-, reduction- and time-parallel computation (sometimes for thread-parallel computation). The second section, contains the cells $C_1, \ldots C_q$, mainly used for thread-parallel computation. Each cell c_i is a minimalist implementation of a processing element or of an execution unit, while each C_j can be a strong and complex processing element. One C_i core is used as ***controller*** (C_1 in our representation) for the MANY-CORE array. In real applications the system is optimal for $p >> q$.

The MANY-CORE section is involved mainly in ***intense computations*** (characterized by: many-core, sequence computing, high-latency functional pipe, buffer-based memory hierarchy), while the MULTI-CORE section is mainly responsible for ***complex computations*** (characterized by: mono/multi-core, multi-threaded programming model, cache-based memory hierarchy) (details in [31]).

The main differences between complex and intense computation at this level of description are: (1) $p >> q$ and (2) the access to the system memory through the cache memory for complex computation and through an explicitly controlled (multi-level) buffer for intense computation.

3) Vertical vs. Horizontal Processing

The buffer-based memory hierarchy allows two kinds of intense computation in the MANY-CORE section. Because the first level **Buffer** module stores m p-element sequences which can be seen as a two-dimension array, the computation can be organized ***horizontally*** or ***vertically***. Let us consider the following m sequences stored in the first level **Buffer**.

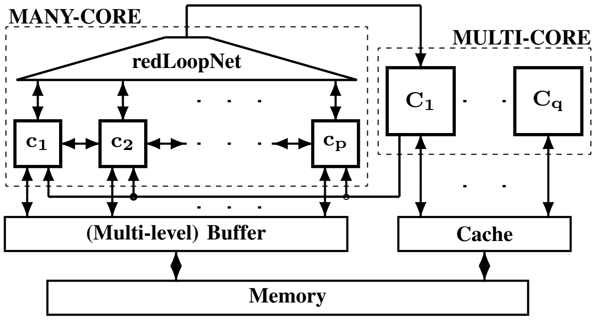


Fig. 5. **The integral parallel abstract machine model.**

$$s_1 =< x_{11}, \ldots, x_{1p} >$$
$$s_2 =< x_{21}, \ldots, x_{2p} >$$
$$\ldots$$
$$s_m =< x_{m1}, \ldots, x_{mp} >$$

Horizontal computing means to consider a function defined on the sequence s_i. For example, FFT on s_i which returns as result the sequences s_{i+1} and s_{i+2}. *Vertical computing* means to compute p times the same function on the sequences

$$< x_{1i}, x_{2i}, \ldots x_{ji} >$$

for $i = 1, 2, \ldots p$. For the same example, p FFT computations can be performed on

$$< x_{1i}, x_{2i}, \ldots x_{ji} >$$

for $i = 1, 2, \ldots p$, with results in

$$< x_{(j+1)i}, x_{(j+2)i}, \ldots x_{(2j)i} >$$

$$< x_{(2j+1)i}, x_{(2j+2)i}, \ldots x_{(3j)i} >$$

for $i = 1, 2, \ldots p$. If $p = j$, the same computation is performed on the same amount of data, but organizing the data for vertical computing has, in this case of the FFT computation, some obvious advantages. Indeed, the "interconnection" between x_{ij} and x_{ik} depends on the value of $|j - k|$ in s_i, while the "interconnection" between x_{ji} and x_{ki} does not depend on the value of $|j - k|$ because the two atoms are processed in the same cell c_i. For other kinds of computations maybe the horizontal computing must be chosen.

The capability of organizing data on two dimensions, vertically or horizontally, allows the use of a p-cell organization to perform computation on data sequences of size different from p. The job to adapt the computation on n-component sequences into a p-cell system organization is left to the compiler. If $n > p$, then the actual sequence will be loaded as few successive p-sized sequences in the two-dimension array $< s_1, \ldots, s_m >$. If $n < p$, then few n-component sequences are accommodated in one p-sized system sequence.

Thus, the MANY-CORE section emulates the computation on a two-dimension network of atoms with a very restrictive but inexpensive horizontal interconnection (due to the linear interconnection between the cells c_i) and a very flexible vertical interconnection (because of the random access in the first level **Buffer**).

C. Recapitulation

A synthetic representation of our abstract model for parallel computation is in Figure 6, where processing is done in the *fine grain* array MANY-CORE and the *coarse grain* array MULTI-CORE.

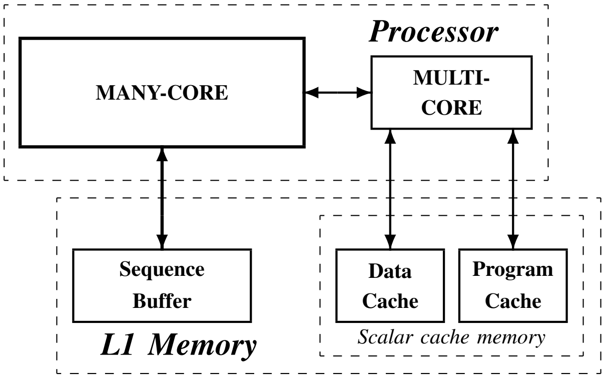


Fig. 6. **Parallel abstract machine model.** Between ***Processor*** and the first level memory hierarchy there are three channels: for program, atoms and sequences.

The first level of memory hierarchy, *L1 Memory*, consists of **Data Cache** for the scalar part of the computation performed mainly by MULTI-CORE, while **Sequence Buffer** is for the sequence computation performed in MANY-CORE. For the code, executed by both, MANY- and MULTI-CORE, there is the **Program Cache**. Due to its high "predictability" the data exchange for the intense computation is supported by a buffer-based memory hierarchy, unlike the complex computation which requests a cache-based memory hierarchy.

The "bottleneck" incriminated by John Backus, for slowing down and making more complex the computation, is not completely avoided. It is substituted only by an enlarged "jarneck", which allows a higher bandwidth between *Processor* and *L1 Memory*. But, while p, the number of cells in MANY-CORE, increases easily from hundreds to thousands, the bandwidth between *L1 Memory* and the system memory is more drastically limited by technological reasons (number of pins, power, technological incompatibilities between logic and DRAMs, ...). The only way to mitigate the effect of this limitation is to design the on-chip **Sequence Buffer** as big as possible in order to avoid frequent data exchange with the off-chip system memory.

There are many possible forms of implementing the abstract model, depending on the targeted application domain. For most of the applications, the use of data-parallel, intense speculative-parallel and reduction-parallel computations covers all the intense computational aspects needed to be accelerated by parallel computation. A good name for this case could be **MapReduce abstract machine model**.

IV. Backus FP Systems as Low Level, Architectural Description

Although Backus's concept of *Functional Programming Systems* (FPS) was introduced as an alternative to the *von Neumann style of programming* in [4], we claim that **they can be seen also as a *low level description* for the parallel computing paradigm**. In the following we use a FPS-like form to provide a low level functional description for the abstract model defined in the previous section. Thus, we obtain the ***virtual machine*** description of a parallel computer, i.e., the description defining the transparent interface between the hardware system and the software system in a real parallel computer. Starting from this virtual machine, the actual ***instruction set architecture*** could be designed for the physical embodiment of various parallel engines.

This section provides, following [4], the low level description for what we call Integral Parallel Machine (IPM). It contains functions which map objects into objects, where an object could be:

- atom, x; special atoms are: T (true), F (false), ϕ (empty sequence)
- sequence of objects, $< x_1, \dots, x_p >$, where x_i are atoms or sequences
- $\perp$: undefined object

The set of functions contains:

- **primitive functions**: the functions performed atomically, which manage:
  - atoms, using functions defined on constant length sequences of atoms, returning constant length sequence of atoms
  - p-length sequences, where p is the number of cells of the MANY-CORE section
- **functional forms** for:
  - expanding to sequences the functions defined on atoms
  - defining new functions
- **definitions**: the programming tool used for developing applications.

A. Primitive Functions

An informal and partial description of a set of primitive functions follows.

- **Atom** : if the argument is an atom, then T is returned, else F is returned.

  $$atom : x \equiv (x\ is\ an\ atom) \rightarrow T; F$$

  The function is performed by the controller or at the level of each c_i cell if the function is applied to each element of a sequence (see *apply to all* in the next subsection).
- **Null** : if the argument is the empty sequence, it returns T, else F.

  $$null : x \equiv (x = \phi) \rightarrow T; F$$

  It is a reduction-parallel function performed by the reduction/loop network, *redLoopNet* (see Figure 5), which returns a predicate to the controller.

- **Equals** : if the argument is a pair of identical objects, then returns T, else F.

  $$eq : x \equiv ((x =< y, z >) \;\&\; (y = z)) \to T; F$$

  If the argument contains two atoms, then the function is performed by the controller, else, if the argument contains two sequences, the function is performed in the cells c_i, and the final results is delivered to the controller through *redLoopNet*.
- **Identity** : is a sort of *no operation* function which returns the argument.

  $$id : x \equiv x$$

- **Length** : returns an atom representing the length of the sequence.

  $$length : x \equiv (x =< x_1, \ldots, x_i >) \to i; (x = \phi) \to 0; \bot$$

  If the sequence is distributed in the MANY-CELL array, then a Boolean sequence, $< b_1, \ldots, b_p >$, with 1 on each position containing a component x_j is generated and *redLoopNet* provides $\sum_1^p b_j$ for the controller.
- **Selector** : if the argument is a sequence with no less than i objects, then the i-th object is returned.

  $$i : x \equiv ((x =< x_1, \ldots, x_p >) \;\&\; (i \leq p)) \to x_i$$

  The function is performed composing an intense speculative-parallel search operation with a data-parallel mask operation and the reduction-parallel OR operation which sends to the controller the selected object.
- **Delete** : if the first argument, k, is a number no bigger than the length of the second argument, then the k-th element in the second argument is deleted.
  $del : x \equiv (x =< k, < x_1, \ldots, x_p >>) \,\&\, (k \leq p) \to$
  $< x_1, \ldots, x_{k-1}, x_{k+1}, \ldots >$
  The *ORprefix* circuit included in the *redLoopNet* subsystem selects the sequence $< x_k, x_{k+1}, \ldots >$, then the left-right connection in the MANY-CELL array is used to perform a one position left shift in the selected subsequence.
- **Insert data** : if the second argument, k, is a number no bigger than the length of the third argument, then the first argument is inserted in the k-th position in the last argument.
  $ins : x \equiv (x =< y, k, < x_1, \ldots, x_p >>) \,\&\, (k \leq p) \to$
  $< x_1, \ldots, x_{k-1}, y, x_k, \ldots >$
  The *ORprefix* function performed in the *redLoopNet* subsystem selects the sequence $< x_k, x_{k+1}, \ldots >$, then the left-right connection in the MANY-CELL array is used to perform one position right shift in the selected sub-sequence and write y in the freed position.
- **Rotate** : if the argument is a sequence, then it is returned rotated one position left.

  $$rot : x \equiv (x =< x_1, \ldots, x_p >) \to < x_2, \ldots, x_p, x_1 >$$

  The *redLoopNet* subsystem and the left-right connection in the MANY-CELL array allows this operation.
- **Transpose** : the argument is a sequence of sequences which can be seen as a two-dimension array. It returns a sequence of sequences which represents the transposition of the argument matrix.
  $trans : x \equiv$
  $(x =<< x_{11}, \ldots, x_{1m} >, \ldots, < x_{n1}, \ldots, x_{nm} >>) \to$
  $<< x_{11}, \ldots, x_{n1} >, \ldots, < x_{1m}, \ldots, x_{nm} >>$
  There are two possible implementations. First, it is naturally solved in the MANY-CELL section because, loading each component of x "horizontally", as a sequence in *Buffer*, we obtain, associated to each cell c_i, the n-component final sequences on the "vertical" dimension (see paragraph 3.2.3):

  $$\begin{array}{c} < x_{11}, \ldots, x_{n1} > \text{ accessed by } c_1 \\ < x_{12}, \ldots, x_{n2} > \text{ accessed by } c_2 \\ \ldots \\ < x_{1m}, \ldots, x_{nm} > \text{ accessed by } c_m \end{array}$$

  where each initial sequence is a m-variable "line" and each final sequence is n-variable "column" in **Buffer**. Second, using rotate and inter sequence operations.
- **Distribute** : returns a sequence of pairs; the i-th element of the returned sequence contains the first argument and the i-th element of the second argument.
  $distr : x \equiv (x =< y, < x_1, \ldots, x_p >>) \to$
  $<< y, x_1 >, \ldots, < y, x_p >>$
  The function is performed in two steps: (1) generates the p-length sequence $< y, \ldots, y >$, then (2) performs $trans << y, \ldots, y >, < x_1, \ldots, x_p >>$.
- **Permute** : the argument is a sequence of two equally length sequences; the first defines the permutation, while the second is submitted to the permutation.
  $perm : x \equiv$
  $(x =<< y_1, \ldots, y_p >, < x_1, \ldots, x_p >>) \to$
  $< x_{y_1}, \ldots, x_{y_p} >$
  With no special hardware support it is performed in time $O(p)$. An optimal implementation, in time belonging to $O(log\, p)$, involves a *redLoopNet* containing a Waksman permutation network, with $< y_1, \ldots, y_p >$ used to program it.
- **Search** : the first argument is the searched object, while the second argument is the target sequence; returns a Boolean sequence with T on each match position.
  $src : x \equiv (x =< y, < x_1, \ldots, x_p >>) \to$
  $< (y = x_i), \ldots, (y = x_p) >$
  It is an intense speculative-parallel operation. The scalar y is issued by the controller and it is searched in each cell generating a Boolean sequence, distributed along the cells c_i in MANY-CELL, with T on each match position and F on the rest.
- **Conditioned search** : the first argument is the searched object, the second argument is the target sequence, while the third argument is a Boolean sequence (usually generated in a previous search or conditioned search); the search is performed only in the positions preceded by T in the Boolean sequence; returns a Boolean sequencer with T on each conditioned match position.
  $csrc : x \equiv$
  $(x =< y, < x_1, \ldots, x_p >, < b_1, \ldots, b_p >>) \to$
  $< c_1, \ldots, c_p >$

where: $c_i = ((y = x_i)\ \&\ b_{i-1})\ ?\ T : F$.
The combination of `src` or `csrc` allows us to define a `sequence_search` operation (an application is described in [32]).

- **Arithmetic & logic operations** :

$$op2 : x \equiv ((x =< y, z >)\ \&\ (y, z\ atoms)) \rightarrow y\, op2\, z$$

where: $op2 \in \{add, sub, mult, eq, lt, gt, leq, and, or, ...\}$
or

$$op1 : x \equiv ((x = y)\ \&\ (y\ atom)) \rightarrow op1\, y$$

where: $op1 \in \{inc, dec, zero, not\}$. These operations will be applied on sequences of any length using the functional forms defined in the next sub-section.

- **Constant** : generates a constant value.

$$\bar{x} : y \equiv x$$

B. Functional Forms

A functional form is made of functions that are applied to objects. They are used to define complex functions, for an IPM, starting from the set of primitive functions.

- **Apply to all** : represents the ***data-parallel*** computation. The same function is applied to all elements of the sequence.

$$\alpha f : x \equiv (x =< x_1, \ldots, x_p >) \rightarrow < f : x_1, \ldots, f : x_p >$$

Example:
$\alpha\, add :<< x_1, y_1 >, \ldots, < x_p, y_p >> \rightarrow$
$< add :< x_1, y_1 >, \ldots, add :< x_p, y_p >>$
expands the function add, defined on atoms, to be applied on sequences, $<< x_1, \ldots, x_p >< y_1, \ldots, y_p >>$, transposed in a sequence of pairs $< x_i, y_i >$.

- **Insert** : represents the ***reduction-parallel*** computation. The function f has as argument a sequence of objects and returns an object. Its recursive form is:
$/f : x \equiv ((x =< x_1, \ldots, x_p >)\, \&\, (p \geq 2)) \rightarrow$
$f :< x_1, /f :< x_2, \ldots, x_p >>$
The resulting action looks like a sequential process executed in $O(p)$ cycles, but on the Integral Parallel Abstract Model (see Figure 5) it is executed as a reduction function in $O(log\, p)$ steps in the *redLoopNet* circuit.
- **Construction** : represents the ***speculative-parallel*** computation. The same argument is used by a sequence of functions.

$$[f_1, \ldots, f_n] : x \equiv < f_1 : x, \ldots, f_n : x >$$

- **Composition** : represents ***time-parallel*** computation if the computation is applied to a stream of objects. By definition:
$(f_q \circ f_{q-1} \circ \ldots \circ f_1) : x \equiv$
$f_q : (f_{q-1} : (f_{q-2} : (\ldots : (f_1 : x) \ldots)))$
The previous form is:
  - sequential computation, if only one object x is considered as input variable
  - pipelined *time-parallel* computation, if a ***stream*** of objects, $|x_n, \ldots, x_1|$, are considered to be inserted, starting with x_1, in c_1 in the MANY-CORE section (see Figure 5) so as in each successive two cells, c_i and c_{i+1}, are performed

$$f_i(f_{i-1} : (f_{i-2} : (\ldots : (f_1 : x_j) \ldots)))$$
$$f_{i+1}(f_i : (f_{i-1} : (\ldots : (f_1 : x_{j-1}) \ldots)))$$

  Thus, the array of cells $c_1, \ldots, c_p$ can be involved to compute in parallel the function

$$f(x) = (f_q \circ f_{q-1} \circ \ldots \circ f_1) : x$$

  for maximum q values of x.

- **Threaded construction** : is a special case of construction for: $f_i = g_i \circ i$ which represents the ***thread-parallel*** computation:
$\theta[f_1, \ldots, f_p] : x \equiv$
$(x =< x_1, \ldots, x_p >) \rightarrow < g_1 : x_1, \ldots, g_p : x_p >$
where: $g_1 : x_1$ represents an independent thread.
- **Condition** : represents a conditioned execution.
$(p \rightarrow f; g) : x \equiv$
$((p : x) = T) \rightarrow f : x; ((p : x) = F) \rightarrow g : x$
- **Binary to unary** : is used to express any function as an unary function.

$$(bu\, f\, x) : y \equiv f :< x, y >$$

This function allows the algebraic manipulation of programs.

C. Definitions

Definitions are used to write programs conceived as functional forms.

Def $new\_function\_symbol \equiv functional\_form$

Example : Let be the following definitions used to compute the *sum of absolute difference* (SAD) of two sequence of numbers:

Def $SAD \equiv (/+) \circ (\alpha ABS) \circ trans$
Def $ABS \equiv lt \rightarrow (sub \circ REV); sub$
Def $REV \equiv (bu\, perm < \bar{2}, \bar{1} >)$

D. Recapitulation

The beauty of the relation between the abstract machine components resulting from Kleene's model and the FPS proposed by Backus is that all the five meaningful forms of composition correspond to the main functional forms, as follows:

Kleene's parallelism	↔	**Backus's functional forms**
data-parallel	↔	`apply to all`
reduction-parallel	↔	`insert`
speculative-parallel	↔	`construction`
time-parallel	↔	`composition`
thread-parallel	↔	`threaded construction`

Let us agree that Kleene's model, and the FPS proposed by Backus represent a solid foundation for parallel computing, avoiding risky *ad hoc* constructs. The generic parallel structure proposed in the next section is a promising start in saving us from saying "Hail Mary" (see [22]) when we decide what to do in order to improve our computing machines with parallel features.

V. A Generic Parallel Engine

The fourth step in trying to restart the parallel computation domain (see subsection I.C) is to propose a simple and, as much as possible, efficient ***generic*** embodiment, able to provide both, a good use of energy and area, and an easy to program computing engine. This section describes the simplest one-chip generic parallel engine, already implemented in silicon: the ***Connex System*** chip. The next four subsections present the organization, the programming style, rough estimates for the 13 Berkeley's motifs, and the physical performances of this first embodiment, which will be, for sure, the subject of many successive improvements.

A. The Organization

The ***Connex System***, centered on the many-cell engine ***ConnexArray***^{TM}, is the first, partial embodiment (see [30]) of the *integral parallel abstract machine model* (see Figure 5). It is proposed as a possible initial version for a generic parallel engine. While the **Cache** module in Figure 5 is by default a hidden physical resource, the *Buffer* module is an explicit part of the architecture that differentiates strongly ***Connex System*** from a standard architecture. Let's start by representing the content of the *Buffer* by $A =< v_1, v_2, \ldots, v_m >$, a two-dimension array containing m p-scalar vectors:

$$\begin{gathered} v_1 =< x_{11}, \ldots, x_{1p} > \\ v_2 =< x_{21}, \ldots, x_{2p} > \\ \ldots \\ v_m =< x_{m1}, \ldots, x_{mp} > \end{gathered}$$

where: each "column", $< x_{1i}, \ldots, x_{mi} >$, is a "vertical" vector of scalars associated to the computational cell c_i in Figure 5. In the first embodiment, represented in Figure 7, **Linear ARRAY of CELLS** is a linear array of p *Connex cells*, $cc_1, \ldots, cc_p$. Each cc_i contains c_i (see Figure 5) and the **local memory** which stores the associated "vertical" vector. The set of local memories represents the first level of **(Multi-level) Buffer** which allows the engine to work as a **stream processor**, not as a simple *vector processor*. Each cell is connected only to the left and to the right cell.

In the ***data-parallel mode*** each Connex cell, cc_i, receives, trough the **Broadcast** net, the same instructions issued by one of the threads, called the *control thread*, running on the **Multi-Threaded Processor**.

The ***reduction-parallel mode*** sends, to the control thread running on the **Multi-Threaded Processor**, scalars returned by functions performed, in the *log*-depth **Reduction** circuit, on sequences of atoms distributed over the array of cells cc_i. The reduction net is a pipelined tree *circuit* because there are only a small number of meaningful reduction functions in the current applications.

In ***speculative-parallel mode*** the difference between h_i and h_j can be done in two ways: (1) by some local parameters specific for each cell (example: its index), or (2) by a specific program loaded in the local memory of each cell. The process is triggered by the *control thread*, while the variable x is issued by the same thread.

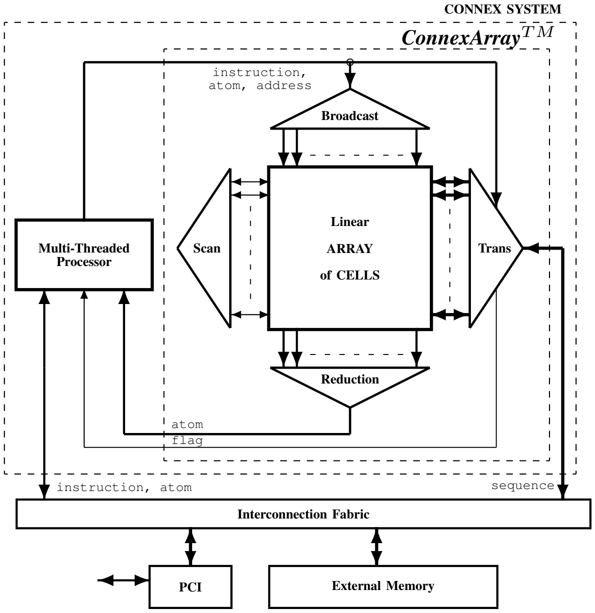


Fig. 7. **The Connex System.** It is centered on ***ConnexArray***^{TM}, a linear array of p execution units, each with its own local memory, connected to the external memory through the **Trans** network. The array has two loops, an external one through **Reduction** net, **Multi-Threaded Processor** and **Broadcast** net, and an internal one through the **Scan** net.

The ***time-parallel mode*** uses the linear interconnection network to configure a pipe of p machines, each controlled by the program stored in the **local memory**. The resulting pipeline machine receives a stream of data and processes it with a latency in $O(p)$. The i-th cell computes function f_i (see subsection 3.1.4). Cell 1 receives rhythmic a new component of the input stream x.

The ***thread-parallel mode*** can be implemented in two ways: (1) each cell works like an independent processor running the locally stored program on local data, or (2) **Multi-Threaded Processor** is used to run, usually, 4 to 16 threads, including the threads used to control the array of cells and the IO process. The second way is more frequently used in the current application domains.

The **Trans** module connects the array to the external memory. It is controlled by one thread running on the **Multi-Threaded Processor**, and works transparent to the computation done in the array.

The global loop closed over the array through **Scan** takes from each cell an atom and computes global functions sending back in the array a sequence of atoms. One example is the function `first` defined on a sequence of Booleans (see subsection 2.2). Another example is the permutation function for which the **Scan** network is programmed with $(-1+2\times log_2 p)$ p-length Boolean sequences.

B. Machine Level Programming for the Generic Parallel Engine

A low level programming environment, called Backus-Connex Parallel FP system – BC for short –, was defined in Scheme for this generic parallel engine (see [19]). Some of the most used functions working on the previously defined array A are listed below:

```
(SetVector a v); a: address, v: vector content
(UnaryOp x)    ; x: scalar|vector
(BinaryOp x y) ; (x,y): scalar | vector
(Cond x y)     ; (x,y): scalar | vector
(RedOp v)      ; RedOp = {RedAdd, RedMax,...}
(ResetActive)  ; activate all cells
(Where b)      ; active where vector b is 1
(ElseWhere)    ; active where vector b was 0
(EndWhere)     ; return to previous active
```

Let us take as example the function *conditioned reduction add*, CRA, which returns the sum of all the components of the sequence $s_1 =< x_{11}, \ldots, x_{1p>}$ corresponding to the positions where the element in the sequence $s_2 =< x_{21}, \ldots, x_{2p} >$ is *less or equal than* the element of the sequence $s_3 =< x_{31}, \ldots, x_{3p} >$:

$$CRA(s_1, s_2, s_3) = \sum_{i=1}^{p} (x_{2i} \leq x_{3i})\,?\,x_{1i} : 0$$

The computation of this function is expressed as follows:

$$\textbf{Def}\ CRA \equiv (/+)\circ(\alpha((leq\circ(bu\ del 1)) \rightarrow (id\circ 1); \bar{0}))\circ trans$$

where the argument must be a sequence of three sequences:

$$x =< s_1, s_2, s_3 >$$

and the result is returned as an atom. For

$$x =<< 1,2,3,4 >, < 5,6,7,8 >, < 8,7,6,5 >>$$

the evaluation is the following:

$CRA : x \Rightarrow$

$(/+) \circ (\alpha((leq \circ (bu\ del\ 1)) \rightarrow (id \circ 1); \bar{0})) \circ trans :$

$<< 1,2,3,4 >, < 5,6,7,8 >, < 8,7,6,5 >>\Rightarrow$

$(/+) \circ (\alpha((leq \circ (bu\ del\ 1)) \rightarrow (id \circ 1); \bar{0})) :<< 1,5,8 >, < 2,6,7 >, < 3,7,6 >< 4,8,5 >>\Rightarrow$

$(/+) :<$

$((leq \circ (bu\ del\ 1)) \rightarrow (id \circ 1); \bar{0}) :< 1,5,8 >,$

$((leq \circ (bu\ del\ 1)) \rightarrow (id \circ 1); \bar{0}) :< 2,6,7 >,$

$((leq \circ (bu\ del\ 1)) \rightarrow (id \circ 1); \bar{0}) :< 2,6,7 >,$

$((leq \circ (bu\ del\ 1)) \rightarrow (id \circ 1); \bar{0}) :< 4,8,5 >>\Rightarrow$

$(/+) :<\_((leq :< 5,8 >) \rightarrow (id : 1); \bar{0}), \ldots, ((leq :< 8,5 >) \rightarrow (id : 4); \bar{0}) >\Rightarrow$

$(/\underline{+}) :< ((leq :< 5,8 >) \rightarrow 1; \bar{0}), \ldots, ((leq :< 8,5 >) \rightarrow 4; \bar{0}) >\Rightarrow$

$(/+) :< (T \rightarrow 1; 0), (T \rightarrow 2; 0), (F \rightarrow 3; 0), (F \rightarrow 4; 0) >\Rightarrow$

$(/+) :< 1,2,0,0 >\Rightarrow 3$

At the level of machine language the previous program is translated into the following BC code:

```
(define (CRA v0 v1 v2 v3)
    (Where (Leq (Vec v2) (Vec v3)))
        (SetVector v0 (Vec v1))
      (ElseWhere)
        (SetVector v0 (MakeAll 0))
    (EndWhere)
    (RedAdd (Vec v0))
)
```

The function `CRA` returns a scalar and has as side effect the updated content of the vector `v0`.

C. Short Comments about Application Domains

The efficiency of ***Connex System*** in performing all the aspects of intense computation remains to be proved. In this subsection we sketch only the complex process of evaluation using the report "A View from Berkeley" [3]. Many decades just an academic topic, "parallelism" becomes an important actor on the market after 2001 when the clock rate race stopped. This research report presents 13 computational motifs which cover the main aspects of parallel computing. Short comments follows about how the proposed architecture and generic parallel engine work for all of the 13 motifs.

For **dense linear algebra** the most used operation is the inner product (IP) of two vectors. It is expressed in FP System as follows:

$$\textbf{Def}\ IP \equiv (/+) \circ (\alpha\times) \circ trans$$

while the BC code is:

```
(define (IP v0 v1)
    (RedAdd (Mult v0 v1))
)
```

allowing a linear acceleration of the computation.

For **sparse linear algebra** the band arrays are first transposed using the function `Trans` in a number of vectors equal with the width w of the band. Then the main operations are naturally performed using the appropriate `RotLeft` and `RotRight` operations. Thus, the multiplication of two band matrices is done on Connex System in $O(w)$.

For **spectral methods** the typical example is FFT. The vertical and horizontal vectors defined in the array A help the programmer to adapt the data representation to obtain an almost linear acceleration [5], because the **Scan** module is designed to hide the performance of the matrix transpose operation. In order to eliminate the slowdown caused by the rotate operations, the stream of samples are operated as vertical vectors (see also [16], where for example: FFT for 1024 floating point samples is done in less than 1 clock cycle per sample).

***N*-Body method** fits perfect on the proposed architecture, because for $j = 0$ to $j = n - 1$ the following equation is computed:

$$U(x_j) = \sum_i F(x_j, X_i)$$

using one cell for each function $F(x_j, X_i)$, followed b the sum (a *reduction* operation).

Structured grids are distributed on the two dimensions of the array A. Each processor is assigned a column of nodes. Each node has to communicate only with a small, constant number of neighbor nodes on the grid, exchanging data at the end of each step. The system works like a cellular automaton.

Unstructured grids problems are updates on an irregular grid, where each grid element is updated from its neighbor grid elements. Parallel computation is disturbed by the non-uniformity of the data distribution. In order to solve the non-uniformity problem a preprocessing step is required to generate an easy manageable representation of the grid. Slowdowns are expected compared with the structured grid.

The typical example of **mapReduce** computation is the Monte Carlo method. This method is highly parallel because it consists in many completely independent computations working on randomly generated data. It requires the add reduction function. The computation is linearly accelerated.

For **combinational logic** a good example is AES encryption which works in 4×4 arrays of bytes. If each array is loaded in one cell, then the processing is pure data-parallel with linear acceleration.

For **graph traversal** in [21] are reported parallel algorithms achieving asymptotically optimal $O(|V| + |E|)$ work complexity. Using sparse linear algebra methods, the breadth-first search for graph traversal is expected to be done on a Connex System in time belonging to $O(|V|)$.

For **dynamic programming** the Viterbi decoding is a typical example. The parallel strategy is to distribute the states among the cells. Each state has its own distinct cell. The inter-cell communication is done in a small neighborhood. Each cell receives the stream of data which is thus submitted to a speculative computation. The work done on each processor is similar. The last stage is performed using the reduction functions. The degree of parallelism is limited to the number of states considered by the algorithm.

Parallel **back-track** is exemplified by the SAT algorithm which runs on a p-cell engine by choosing $log_2\ p$ literals, instead of one on a sequential machine, and assigning for them all the values from $00\ldots0$ to $11\ldots1 = p - 1$. Each cell evaluates the formula for one value. For parallel **branch & bound** we use the case of the Quadratic Assignment Problem. The problem deals with two $N \times N$ matrices: $A = (a_{ij})$, $B = (b_{kl})$. The global cost function:

$$C(p) = \sum_{i}^{n} \sum_{j}^{n} a_{ij} \times b_{p(i)p(j)}$$

must be minimized finding the permutation p of the set $N = \{1, 2, \ldots, n\}$. Dense linear algebra methods, efficiently running on our architecture, are involved here.

Graphical models are well represented by parallel hidden Markov models. The architectural features reported in research papers refers to fine-grained data-parallel processor arrays connected to each node of a coarse-grained PC-cluster. Thus, our engine can be used efficiently as an accelerator for general purpose sequential engines.

For **finite state machine** (FSM) the authors of [3] claim that ”nothing helps”. But, we consider that the array of cells with their local memory loaded with non-deterministic FSM descriptions work very efficient as a speculative engine for applications such as deep packet inspection, for example.

At the end of this superficial introductory analysis, *which must be deepened by future investigations*, we claim that for almost all the computational motifs the ***Connex System***, in its simple generic form, perform at least encouraging if not pretty well.

D. About the First Implementation

Actual versions of the Connex System have already been implemented as part of a SoC designed for HDTV market: *BA1024* (see [28], [29] and [31]). The $90\,nm$ version of *BA1024*, with 1024 16-bit EUs, is in Figure 8. The last version, implemented in $65\,nm$, provides a peak performance of $400\,GOPS$[6], which translates in:

- area efficiency: $> 6\,GOPS/mm^2$
- power efficiency: $> \mathbf{120\,GOPS/Watt}$

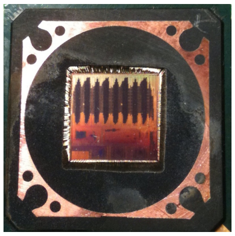


Fig. 8. **The Connex Chip.** The $90\,nm$ version of *BA1024* chip. The Connex System uses 60% of the total area.

Compared with a standard sequential processor implemented in $65\,nm$ results $20\times$ in area use and $100\times$ in power use (the evaluation was made for applications involving 32-bit integer and float operations). For integer (no-float) applications the previous improvements are around 4 times higher (the actual measurements were made for programs running HDTV *frame rate conversion* on the $65\,nm$ version of the *BA1024* chip). This first implementation of a generic parallel system suggests that *genuine parallelism is naturally green*.

The performances of this first embodiment of a *generic parallel structure* looks so good because the architecture is focused on intense computation, starting from the idea, largely exposed in [31], that only the strong ***segregation between intense computation and complex computation*** allows a good use of *area and power*. The standard sequential processors perform on the same hardware both, complex and intense computation, but they are designed and optimized only for complex computation. More, the multi-core systems are *ad-hoc* constructs gathering together few standard sequential processors able to perform efficiently no more than complex multi-threaded computations. Many-core GPU systems, like ATI or NVIDIA, do not obey to the golden rule “small is beautiful” stated in [3] (see pag. 20), to which we must add that “simple is efficient”. Thus, the main GPUs do not obey (our version of) the “*kiss* principle”: *keep it small & simple*. Unlike the Connex approach, they have complex hardware (hardware multipliers, float units), cache memories (why caches for a very predictable flow of data and programs!?), hierarchical organization (unrequested by any computational model), most of them imposed unfortunately by oppressive legacies.

VI. Conclusion

The intrinsic parallel computational model of Kleene fits perfect as theoretical foundation for parallel computation. Because, as we proved, primitive recursion and minimalization

[6] GOPS: **G**iga 16-bit **O**perations **P**er **S**econd

are particular forms of compositions, only the composition rule is used to highlight the five forms of parallelism: data-, speculative-, reduction-, time-, thread-parallelism.

Integral Parallel Abstract Machine Model is defined as the model of the simplest generic parallel engine. Real embedded computation frequently involves all forms of parallelism for running efficiently complex applications.

Both, Kleene's model and Backus's programming style promote one-dimension arrays, thus supporting the simplest hardware configuration for the initial architectural proposal. If needed, a more complex solution will be adopted. But, till then we must struggle in this initial context inventing new algorithms by keeping the hardware as *small & simple* as possible. The linear array of cells is complemented by four *log*-depth networks – Broadcast, Reduction, Trans and Scan – in order to compensate its simplicity.

Segregation between intense computation and complex computation is the golden rule for increasing the computational power lowering in the same time both, cost (silicon area) and energy.

Parallelism is meaningful only if it is "green". The one-chip solution for parallelism provides a low power computational engine if it is developed starting from genuine computational and abstract models. The Connex Chip proves that our proposal for a generic one-chip solution provides two magnitude orders improvement in reducing energy per computational task compared with standard sequential computation, while GPGPU-like chips, an *ad hoc* solution for parallelism, are unable to provide the expected reduction of energy per computational task (they consume hundreds of Watts per chip and are unable to achieve, on average, more than 25% of their own peak performance running real applications [1]).

Programming the generic parallel structure is simple because of the simplicity of its organization, easy to hide behind a well defined architecture. Backus's FP Systems capture naturally the main features of a parallel engine and provide a flexible environment for designing an appropriate generic parallel architecture. In this paper we didn't touch the problem of a programming model, because it must be based, in our opinion, on the insights provided in the fifth stage of our approach – the algorithmic evaluation of the generic parallel structure against the 13 Berkeley's computational motifs (see 1.3). All the current programming models (such as Intel TBB, Java Concurrency, .Net Task Parallel Library, OpenMP, Clik++, CnC, X10) are basically focused on the multi-threading computation because the current market provides the multi-core hardware support for this kind of parallelism, while the scalable programming models presented in [20] are focused on the current many-core market products (such as AMD/ATI, NVIDIA).

Future work refers to the last stage of our approach (see subsection I.C). Preliminary evaluations tell us that almost all the 13 computational motifs, highlighted by *Berkeley's view* in [3], are reasonable well supported by the proposed generic parallel structure initially programmed in a sort of FP System programming language (for example BC). During this evaluation a lot of new features will be added to the simple generic engine described in the present paper. The resulting improvement process will allow a gradual maturation of the concept of parallelism, because *Nihil simul inventum est et perfectum* (Marcus Tullius Cicero).

Acknowledgment

The authors got a lot of support from the main technical contributors to the development of the *ConnexArray*^{TM} technology, the *BA1024* chip, the associated language, and its first application: Emanuele Altieri, Frank Ho, Bogdan Mîţu, Marius Stoian, Dominique Thiebaut, Tom Thomson, Dan Tomescu. The comments received from Bogdan Niţulescu helped a lot the improvement of the paper.

References

[1] ***, http://www.siliconmechanics.com/files/c2050benchmarks.pdf. Silicon Mechanics, 2012.

[2] B. Alpern, L. Carter, and J. Ferrante, Modeling parallel computers as memory hierarchies. In Giloi, W. K. *et al.* eds. *Programming Models for Massively Parallel Computers*, IEEE Press, 1993.

[3] K. Asanovic, *et al.*, The landscape of parallel computing research: A view from Berkeley, 2006. At: http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.pdf.

[4] J. Backus, Can programming be liberated from the von Neumann style? A functional style and its algebra of programs. *Communications of the ACM* 21, 8 (August) 1978. 613641.

[5] C. Bîra, L. Gugu, M. Maliţa, G. M. Ştefan: Maximizing the SIMD Behavior in SPMD Engines, in *Proceedings of the World Congress on Engineering and Computer Science 2013 Vol I WCECS 2013*, 23-25 October, 2013, San Francisco, USA. 156-161.

[6] G. Blaauw, and F.P. Brooks, The structure of System/360, part I - Outline of the logical structure. IBM Systems Journal 3, 2, 1964. 119135.

[7] A. Church, An unsolvable problem of elementary number theory. *The American Journal of Mathematics* 58, 1936. 345363.

[8] D. Culler, *et al.*, LogP: Toward a realistic model of parallel computation. *Proc. of the ACM SIGPLAN Symposium on Principles and Practices of Parallel Programming*, 1991. 112.

[9] M. Davis, *The Undecidable. Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions*. Dover Publications, Inc., Mineola, New-York, 2004.

[10] E. W. Dijkstra, Co-operating sequential processes. *Programming Languages* Academic Press, New York, 43112. Reprinted from: Technical Report EWD-123, Technological University, Eindhoven, the Netherlands, 1965.

[11] S. Fortune, and J. C. Wyllie, Parallelism in random access machines. *Conference Record of the Tenth Annual ACM Symposium on Theory of Computing*, 1978. 114118.

[12] L. M. Goldschlage, A universal interconnection pattern for parallel computers. *Journal of the ACM* 29, 4, 1982. 10731086.

[13] R. Greenlaw, H. J. Hoover, and W. L. Ruzzo, *Limits to Parallel Computation*. Oxford University Press, 1995.

[14] S. Kleene, General recursive functions of natural numbers. *Mathematische Annalen* 112, 5, 1936. 727742.

[15] R. E. Ladner, and M. J. Fischer, Parallel prefix computation. *Journal of the ACM* 27, 4, 1980. 831838.

[16] I. Lorentz, M. Maliţa, R. Andonie Fitting fft onto an energy efficient massively parallel architecture. *The Second International Forum on Next Generation Multicore / Manycore Technologies*, 2010

[17] B. M. Maggs, L. R. Matheson, and R. E. Tarjan, Models of parallel computation: a survey and synthesis. *Proceedings of the Twenty-Eighth Hawaii International Conference on System Sciences*, 2, 1995. 6170.

[18] M. Maliţa, and G. Ştefan, On the many-processor paradigm. *Proceedings of the 2008 World Congress in Computer Science, Computer Engineering and Applied Computing*, Las Vegas, vol. PDPTA08, 2008. 548554.

[19] M. Maliţa, and G. Ştefan, Backus language for functional nano-devices. *CAS 2011, vol. 2*, 331334.

[20] M. D. McCool, Scalable programming models for massively multicore processors. *Proceedings of the IEEE 96*, 5 (May), 2008. pp. 816831.

[21] D. Merrill, M. Garland, and A. Grimshaw, High Performance and Scalable GPU Graph Traversal, *Technical Report CS-2011-05*, Department of Computer Science, University of Virginia, Aug, 2011.

[22] D. Patterson, The trouble with multicore. *IEEE Spectrum*, July 1, 2010.

[23] E. Post, Finite combinatory processes. formulation I. The Journal of Symbolic Logic 1, 1936. 103105.

[24] V. R. Pratt, M. O. Rabin, and L. J. Stockmeyer, A characterization of the power of vector machines. *Proceedings of STOC1974*, 1974. 122134.

[25] V. R. Pratt, and L. J. Stockmeyer, A characterization of the power of vector machines. *Journal of Computer and System Sciences* 12, 2 (April), 1976. 198221.

[26] V. Sakar, *Parallel computation model*, 2008. At: http://www.cs.rice.edu/vs3/comp422/lecture-notes/comp422-lec20-s08-v1.pdf.

[27] G. Ştefan, and M. Maliţa, Granularity and complexity in parallel systems. *Proceedings of the 15 IASTED International Conf*, 2004. 442447.

[28] G. Ştefan, The CA1024: A massively parallel processor for cost-effective HDTV. *Spring Processor Forum: Power-Efficient Design*, May 15-17, Doubletree Hotel, San Jose, CA.

[29] G. Ştefan, *et al.*, The CA1024: A fully programmable system-on-chip for cost-effective HDTV media processing. *Hot Chips: A Symposium on High Performance Chips*. Memorial Auditorium, Stanford University.

[30] G. Ştefan, Integral parallel architecture in system-on-chip designs. *The 6th International Workshop on Unique Chips and Systems*, Atlanta, GA, USA, December 4, 2010, pp. 2326.

[31] G. Ştefan, One-chip TeraArchitecture. *Proceedings of the 8th Applications and Principles of Information Science Conference*. Okinawa, Japan, 2009.

[32] D. Thiebaut and M. Maliţa, "Real-time Packet Filtering with the Connex Array" *The 33rd Annual International Symposium on Computer Architecture*, Boston, MA, USA June 17-21, 2006, pp. 17-21.

[33] A. M. Turing, On computable numbers with an application to the Eintscheidungsproblem. *Proceedings of the London Mathematical Society* 42, 1936.

[34] J. von Neumann, First draft of a report on the EDVAC. *IEEE Annals of the History of Computing* 5, 4, 1993.

[35] L. G. Valiant, A bridging model for parallel computation. *Communications of the ACM* 33, 8 (Aug.), 1990. 103111.

Gheorghe M. Ştefan teaches digital design in *Politehnica* University of Bucharest. Its scientific interests are focused on digital circuits, computer architecture and parallel computation. In the 1980s, he led a team which designed and implemented the Lisp machine DIALISP. In 2003-2009 he worked as Chief Scientist and co-founder in *Brightscale*, a Silicon Valley start-up which developed the BA1024, a many-core chip for the HDTV market. More at `http://arh.pub.ro/gstefan/`.

Mihaela Maliţa teaches computer science at Saint Anselm College, US. Her interests are programming languages, computer graphics, and parallel algorithms. She wrote and tested different simulators for the Connex parallel chip. More at `http://www.anselm.edu/mmalita`.

An intrusion detection approach using fuzzy logic for RFID system

Ali Razm, Seyed Enayatallah Alavi

Abstract— Fuzzy systems have demonstrated their ability to solve different kinds of problems in various applications domains. Currently, there is an increasing interest to augment fuzzy systems with learning and adaptation capabilities. Two of the most successful approaches to hybridize fuzzy systems with learning and adaptation methods have been made in the realm of soft computing. Neural fuzzy systems and genetic fuzzy systems hybridize the approximate reasoning method of fuzzy systems with the learning capabilities of RFID and evolutionary algorithms. The objective of this paper is to describe a fuzzy genetics-based learning algorithm and discuss its usage to detect intrusion in RFID. For the purpose of training and evaluation of our proposed approach, part of the RFID system-generated dataset provided by the University of Tasmania's School of Computing and Information Systems was used, in addition to simulated datasets. which have information on RFID, during normal behaviour and intrusive behaviour. This paper presents some results and reports the performance of generated fuzzy rules in detecting intrusion in RFID.

Keywords— Intrusion detection; Fuzzy logic; Genetic algorithm; Rule learning.

I. Introduction

OBJECT identity cloning is a serious issue due to its global social and economic impact. According to Aberdeen research group [1], worldwide identity theft losses were $73.8 billion in 2002, reaching $2 trillion by the end of the year 2005 [2], a number expected to increase in the years to come.

According to a report by the World Health Organization (WHO), the number of counterfeit products entering global markets has been increasing every year [3]. The number of counterfeiting incidents in 2007 has increased by ten-fold than it was in the year 2000.

Counterfeiting of medical products imposes serious health risks or even death to patients and causes high economic losses to the industries producing genuine products.

RFID is a promising technology used to combat object identity cloning in applications such as pharmaceutical supply chains because it increases track and trace visibility [4] [5]. RFID is also being increasingly used in various applications such as automatic payment, where high security is required.

For an RFID system to be successfully employed in a ubiquitous environment, the cost of the tag is one of the major factors to consider [6]. RFID tags have to be low cost devices to be viable in many applications. Some estimates suggest that with a large scale production, the cost of EPC tag can drop below five cents. In 2001, researchers at AUTO-ID lab proposed the design for five cent RFID tag, which they pointed it to be difficult but achievable without the need to scale the production of the tag [7]. There are also some successes in producing printed RFID tags, which is estimated to reduce the current cost of silicon-based tags by more than ten-fold [8].

However, due to the trade-off between the cost and computational capability, securing Low-Cost RFID systems has been a remarkable challenge in the research community. This is mainly because reduction in tag storage and processing capacity limits the ability of the tag to provide enough resources for execution of computationally intensive but more secure cryptographic algorithms.

It has been indicated that implementing intrusion detection systems in RFID (at the middleware and backend level) is a promising approach to serve as an additional layer of security on top of the existing security protocols. For example, Mirowski et al. [9] implemented statistical intrusion detection mechanism introduced by Denning [10], while Lehtonen et al.[11] is based on the probabilistic technique.

Most of these approaches achieve high detection rate, however, they come with high false positive rate.

We started by assessing the possibility of implementing an intrusion detection mechanism, based on fuzzy logic, that achieves high detection and low false positive rates on RFID system.

Ali Razm, Department of IT, Electricity Distribution Company of ahvaz, Pasdaran Blvd., Ahvaz, Iran(e-mail: ali.razm23@gmail.com)

Seyed Enayatallah Alavi, Department of Computer Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran (e-mail: se_alavi@yahoo.co.uk).

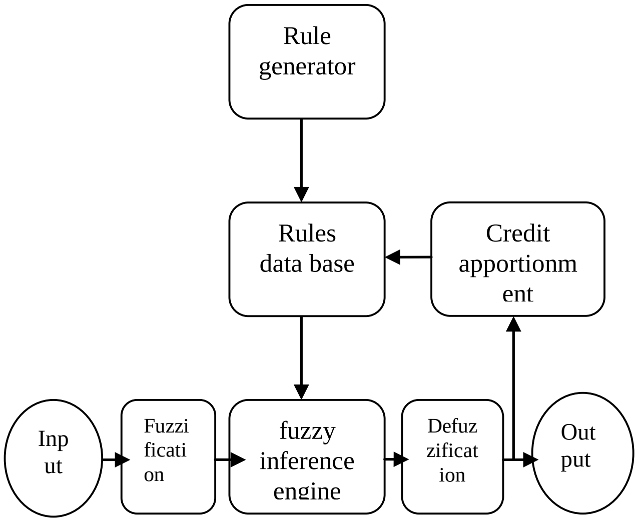


Fig. 1 Hybrid Fuzzy Genetics-Based RFID Anomaly Intrusion Detection System Diagram

II. RELATED WORK

Much of the RFID research over the past few years has been aimed at addressing security and privacy challenges of low-cost tags. For example, Dimitriou [12] proposed a lightweight authentication protocol in order to protect privacy of users and defend tags against cloning. In this protocol, both readers and tags authenticate each other based on shared secret key, which is refreshed every time the tag is read.

Generally, there are three approaches taken to solve security and privacy issues in RFID system. These are:

Low-cost tag hardware design – involves designing a small size, low-cost and computationally efficient tag hardware manufacturing

Lightweight cryptographic algorithm design – involves designing secure algorithms that can run over a low computational capacity tags

Intrusion detection system design – implements mechanisms to detection attacks when they occur in the system

The first approach is the one related to manufacturing of small but efficient hardware technology that can support computationally intensive cryptographic algorithms on tags. This is the area related to re-engineering of the efficient and low-cost tag memory and battery systems.

The second research area is about designing efficient cryptographic algorithms that require less resource for execution. This revolves around designing light weight cryptographic algorithms.

The third approach deals with whether it is possible to come up with efficient additional algorithms on the top of the existing low-cost authentication protocols. This provides additional detection layer, without the need to depend on the tag resources, by avoiding the implementation of complex security protocols on the tag rather on the other components (reader, middleware or the backend system).

III. INTRUSION DATASET

The dataset used for the evaluation of the proposed approach is provided by the University of Tasmania's School of Computing and Information Systems. This dataset has been used by Mirowski et al. [9] and Lehtonen et al. [13] to evaluate cloning detection approaches in RFID systems.

The dataset consists of RFID read events, collected from proximity card system, over the period of three years. The events are generated by the system when the cardholder attempts to unlock the doorways of the building.

Since the dataset contains, information such as: "which user", "which doorways of the building", "at what time of the day", "how often", it can be used to characterize the normal behavior of each user accessing restricted areas of the building. These in turn would help to characterize the abnormal cases where the observation varies from the expected normal.

The proximity card system consists of: the proximity card (RFID tag), the proximity card reader (RFID reader), and the database system at the backend. There are originally six attributes associated to each read event in the database at the backend. Table 1 shows the attributes in the original dataset and their possible values.

Table.1 The Original dataset Attributes and their possible values [14]

Attributes	Description
Date	The date of a year, dd/mm/yyyy
Time	The time 12 hour format, hh:mm:ss
Time Period	AM, PM
Access/Decision	Granted Denied-noPermission Denied-foreignCard Denied-TimeZone
Reader Number	Readers numbered 25 to 28
Card Number	Tag ID number converted to integer values (100 – 413)

A. Tag Cloning

Cloning attack is one of the major concerns in RFID system. Tag cloning occurs when the attacker makes a duplicate of the legitimate tag by being able to gain access to the ID and other data of the legitimate tags. The cloned tag will then be used by the attacker to gain an authorized access to a restricted area or launch any other malicious action by making they appear as the genuine tags. Table 2 summarizes the common types of RFID attacks and interfaces that can be exploited to launch them.

Table.2 RFID attack Types and Exploited Interfaces, Modified from [15]

Attack type	Interface
Eavesdropping	Communication Channel, Tag, Backend
Denial of Service	Communication Channel
Unauthorized Reading	Communication Channel, Tag
Tag Data Modification	Tag
Relay	Communication Channel
Replay	Communication Channel
Cloning	Tag
Malicious SQL injection	Middleware, Backend

B. Denial of Service

Denial of Service (DoS) attack occurs when the attacker uses a blocker tag or jams the normal communication between the legit reader and the tag. During blocking attack, the attacker uses fake tags and simulated the existence of several in the system there by making the reader to endlessly probe the several non existing tags, hence denying service to the legitimate tags.

In jumping attack, the attacker induces large amount of radio noises with the same frequency within the system causing the system to freeze hence successful launching of denial of service.

However, denial of service such as blocking has been proposed in some literatures to be techniques to preserve the privacy of the tag bearer [16].

C. Eavesdropping

Eavesdropping involves unauthorized listening of communication between two or more partieswithout the knowledge of the participants. Eavesdropping can take place in two ways, active and passive. In active eavesdropping, the attacker probes the victim, which in most cases is the tag, and collects the response. In passive eavesdropping, the attacker does not do anything but listens and captures communication between the tag and reader. Passive eavesdroppers are difficult to detect as they are not involved in a communication, but are only listening to the channel.

Passive eavesdropping is prevented in traditional wireless systems by implementing strong encryption algorithms, which are computationally infeasible for the eavesdropper to be able to decrypt it.

D.Replay Attack

In Replay attack, the attacker first eavesdrop the communication between the genuine tag and reader and sends the same procedure, e.g. authentication sequence, to the genuine devices at latter times in an attempt to make system resources unavailable to the legit users (DoS attack).

To detect and protect replay attack, different countermeasure procedures have been given and most of them are based adopting challenge response protocols.

IV. FUZZY GENETICS-BASED MACHINE LEARNING ALGORITHM

A Hybrid fuzzy genetics-based machine learning algorithm combines two genetics-based machine learning algorithms, known as Pittsburgh and Michigan approaches, for designing classification system that makes use of fuzzy rulebase [17]. It makes use of the benefits of both its constituent algorithms to come up with the classification system with better performance.

Pittsburgh approach considers a set of fuzzy rules as an individual and each fuzzy rule set in a population is evaluated by its overall classification performance over the training samples. The fitness of individual rule set is its classification performance (accuracy) over the training samples. During genetic operation, crossover takes place between two fittest individuals (rule sets) while mutation is applied on an individual rule set by modifying the rule set’s elements (rules). Each fuzzy rule sets in Pittsburgh approach is coded as a binary string where each bit indicates the presence or absence of constituent fuzzy rule within the population.

Michigan approach considers each rule within the population as an individual and each fuzzy rule are coded by series of integer/character. The fitness of individual fuzzy rule is calculated by its classification performance over all the training samples. Genetic operations are performed on fuzzy rules in the population where crossover takes place between two best individual fuzzy rules and mutation is performed by modifying constituent antecedent fuzzy sets of the fuzzy rules depending on the given probability.

The advantage of Michigan approach over the Pittsburgh approach is that it has global search ability for a best fuzzy rule over the training samples. Its weakness is that it has low optimization ability of the rulebase used for the system. On the other hand, Pittsburgh approach has ability to optimize rule-base used for classification system. However, Pittsburgh approach has lower global search ability of fuzzy rules over the training samples.

The hybrid approach makes use of the advantage of both approaches. Genetic operators for generating new fuzzy rules are performed by using Michigan approach to heuristically search for the best fuzzy rules while the entire algorithm is based on Pittsburgh approach. Hybrid approach has been shown to have better performance ability than its constituents, Pittsburgh and Michigan approaches [17].

A. Initial Fuzzy Rule-Base Generation

Fuzzy rule bases are normally constructed by using knowledge from human experts who know a domain for which the expert system is being designed very well. Obtaining knowledge from experts, however, is tedious and expensive or can even be erroneous when designing complex system where the amount of information required is too large. To overcome this challenge, various methods have been proposed to automatically learn fuzzy rules from data [18-19-20]. For generating fuzzy rules from numerical data in this paper, we utilize the method proposed by [21] for classification problems.

B. Antecedent Fuzzy set Design

For the classification problem, Nozaki et al. [20] divides the process of constructing fuzzy rules from training data into two phases. In the first phase a pattern space is partitioned into fuzzy subspaces, as illustrated in Figure 2. Partitioning of a pattern space into fuzzy spaces is done in so that it does not result in too few or too many fuzzy partitions. Too fine and too coarse partitioning of a pattern spaces are both undesirable, as they would lead to the poor performance of the model. If the partition is too fine, the model would not generate rules for some of the partitions due to the lack of training data within them. If the partition is too coarse, then the model will generate few rules which are too general leading the model to poor performance.

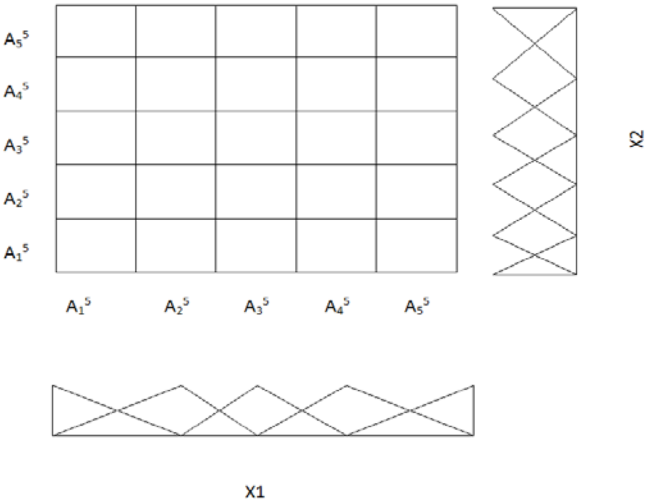


Fig.2 Fuzzy partitioning of the two dimensional pattern space [20]

In the Figure 2, we have fuzzy sets 5×5 which equals 25 fuzzy partitions on two dimensional spaces. The superscript K, in an antecedent fuzzy set A_i^K, represents the maximum number of subspace in each axis and i represents the current partition on the given axis.

Nozaki et al. [21] proposes two possible approaches to avoid the problem of having too few or too many fuzzy partitions. One of the methods considers applying heuristics technique based on the density of the dataset. The other method makes use of multiple fuzzy partitions simultaneously. By simultaneously utilizing multiple fuzzy partitions, the model considers all possibilities of partitioning the pattern space. This is illustrated in Figure 3. The latter approach has been used by Nozaki et al. [20] for pattern classification problem. In this paper, we make use of the latter approach where multiple fuzzy partitions are simultaneously utilized.

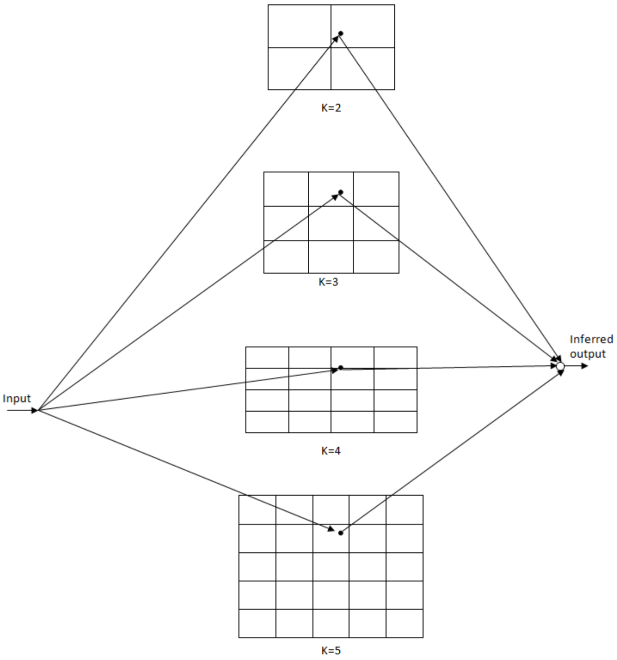


Fig.3 Simultaneous use of multiple pattern spaces for two dimensional patterns [20]

The approach that simultaneously utilizes multiple fuzzy partitions also has some drawbacks when working with high dimensional pattern classification problems. This is because the number of the generated fuzzy rules could be quite large leading to more complex model.

Once the pattern space is partitioned into subspaces and each fuzzy subspace is defined, the next step is to use the combination of each attributes as an antecedent part of the fuzzy rule. For example, in Figure 3, we have $(14+1)^2$ combinations of the fuzzy sets as an antecedent part of the fuzzy rule. The total number of partitions for each attributes is 14 and 1 is for the “don’t care” character. The superscript 2 is the number of attribute in a pattern (the dimension of the pattern classification space). In the partition above (see Figure 3), if there are n-dimensions in the problem, the total number of the generated fuzzy IF-THEN rules are 15^n.

For a low dimensional problem with a few attributes, the number of generated rules is reasonably small and they can all be used as candidate rules. On the other hand, for high dimensional pattern classification problem (large value of n), it is infeasible to examine all the generated rules [23]. To select reasonable number of rules for the pattern classification problem with high dimension, in some approaches rules are randomly selected [24] while some other approaches use heuristic rule evaluation criteria to choose a smaller number of rules as candidate fuzzy rules [23]. For the experiment presented in this paper, there are only three attributes and it is reasonable to consider all the generated fuzzy rules as the candidate rules.

The consequent class for each fuzzy IF-THEN rule (as discussed in next sub-section) is determined by taking the majority class in the corresponding fuzzy subspace. The method is proposed in [21]. The same approach in used for the experiments in this paper.

The given axis the pattern space is divided into K fuzzy sets $\{\alpha_1^k, \alpha_2^k, \ldots, \alpha_k^k\}$. The membership function selected is triangular.

$$f_i^k(x) = \mathrm{Max}\left\{1 - \frac{\left|x - a_i^k\right|}{b^k}, 0\right\}, \qquad i = 1,2,\ldots,k \tag{1}$$

a_i^k of the symmetric triangular membership function in the equation (1) is computed as:

$$a_i^k = (i-1)/(k-1), \quad i=1,2,\ldots,k. \tag{2}$$

b^k in equation(1), representing the spread of the membership function is computed as:

$$b^k = 1/(1-k) \tag{3}$$

C. Determining the Consequent Classes and the Grade of Certainty

The consequent class of the given rule (antecedent) is determined by first calculating the total compatibility grade of the fuzzy rule with each training pattern and assigning the class with the maximum compatibility grade as the consequent class. The following procedure explains how to determine the compatibility grade fuzzy rule with each training pattern. Then the method for computing grade of certainty is given.

There are two operators used to calculate the compatibility grade of fuzzy rules with each training patterns, product and minimum operators. The product operator is more popular while the minimum operator is not as famous [24]. In this paper, we use the product operator to calculate the compatibility grade of the training pattern with antecedent fuzzy sets.

$$\alpha_q(X_p) = \alpha_{q1}(x_{p1}).\alpha_{q2}(x_{p2})...\alpha_{qn}(x_{pn}) \qquad (4)$$

Since our patterns have only three attributes, the compatibility grade will be

$$\alpha_q(X_p) = \alpha_{q1}(x_{p1})\,.\alpha_{q2}(x_{p2})\,.\alpha_{q3}(x_{p3})$$

αq(Xp) is the compatibility of Xp with the antecedent part $\alpha q=(\alpha q1,\alpha q2,\ldots,\alpha q^{i})$ and αq1(.) is the membership function of the antecedent fuzzy part.

The total compatibility grade of each class with antecedent vector is calculated as follows:

$$\beta_{c1} = \sum\nolimits_{x_p \in c1} \alpha_q(X_p) \qquad (5)$$

And

$$\beta_{c2} = \sum\nolimits_{x_p \in c2} \alpha_q(X_p) \qquad (6)$$

If $\beta_{c1} = \beta_{c2}$ then the fuzzy sets corresponding to the fuzzy subspace (α_i^k, α_j^k) is not generated.

For $\beta_{c1} \neq \beta_{c2}$, the class of the partition space will be the majority class in that subspace.

$$\beta_{c1} > \beta_{c2} \qquad \text{then } C_{ij}^{k} = C1 \qquad (7)$$

And

$$\beta_{c1} < \beta_{c2} \qquad \text{then } C_{ij}^{k} = C2 \qquad (8)$$

The the grade of certainty for each rule is calculated as follows:

$$CFijk = \left|\beta_{c1} - \beta_{c2}\right| / (\beta_{c1} + \beta_{c2}) \qquad (9)$$

From the computations in equations 5 and 6, the class with the larger total compatibility grade to the premises of the fuzzy rule set is considered to be the consequent class. The value of the certainty grade is between the interval [0, 1].

From the above procedure, large number of fuzzy rules can be extracted by assigning the consequent class and the rule weight for every possible combination of the antecedent fuzzy sets. However, all these large number of fuzzy rules are intractable for human users. In addition, long fuzzy rules with many antecedent fuzzy rule conditions are difficult for humans to understand. In other words, the generated rule by using only the above procedure suffers from low interpretability. To overcome this, only short fuzzy rules with limited length of antecedent part are normally used. In our experiments, since we only have three attribute, we use them as they are.

V. Experiments

For our experiment, we used events generated by randomly selected users of the proximity cardsystem. The users are selected such that the number of events generated by them is beyond the certain minimum (120 in our case), as users with insufficient records cannot be properly characterized. The choice of 120 as the bottom-line is done by intuition; however, selecting users with larger number of records would provide better result.

We added frequencies of identical events observed in the system within a day, as a new attribute, by computing it from the original dataset and the day attribute. A day (24 hours) and the user's access frequency attributes are normalized to the range [0, 1] by using the MIN-MAX normalization method. Each day of the week (Sunday to Saturday) was assigned a numeric value 1 to 7 respectively.

Every user of the premises is characterized by the frequency of their normal access within a given range of time of a day. Whenever there are events generated by an unauthorized user(s) by obtaining the clone of one or more proximity cards, the total number of events generated by the same ID will raise above normal (the aggregate of the normal user and the non-legitimate user), signaling the presence of anomaly within the system.

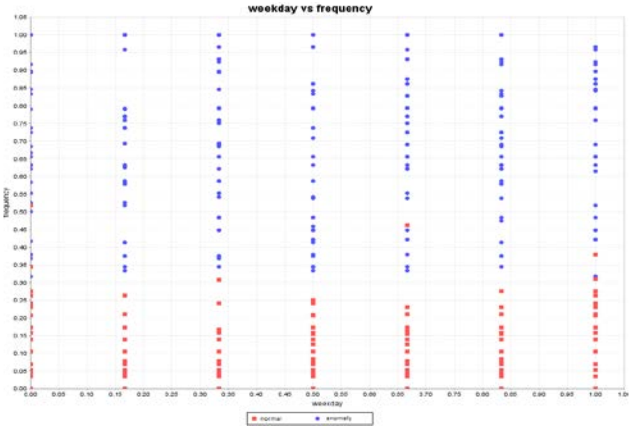


Fig.4 Events generated by the normal users and events generated in the presence an Intruder User(s)

Anomalous data of a given day for a given user (See Figure 4) was generated by using the following formula:

$$T(u,d) = fmax(u, d)+randb(1, 3^* fmax(u,d))$$

Where

T (u,d): is anomalous data of the give user at a given day

fmax(u, d): is the maximum use frequency of the given user at the given

randb(1, 3* fmax(u,d)): is the random number between 1 and thrice the maximum frequency at a given day.

As discussed in the section above, the major limitation of the real world dataset is the fact that it is assumed to be attack-free. In other words, all the obtained events are taken to be generated by normal tags. Thus, synthesizing events to characterize attack behavior was the necessary step inour experiment.

A. Hybrid Fuzzy GBML Parameter Settings

Hybrid Fuzzy Genetics-based Algorithm was applied to the candidate rule set population containing Npop=200 fuzzy rule sets where each fuzzy rule set consists of 10 rules. The probability for the Michigan iteration is set to 0.5 while crossover probability for both Pittsburgh (Table 3).

Table.3 The Hybrid Fuzzy-GBML parameters used

Parameter	**Value**
Number of Fuzzy Rules	20
Number of Rule Sets	200
Crossover probability	0/9
Number of Generations	1000
Do not Care Probability	0/5
Probability for a Michigan Iteration	0/5

The linguistic variables used are: 1 is small, 2 is medium small, 3 is medium, 4 is medium large, 5 is large and # is don't care.

B. Results

The following are the first 10 rules generated after the training of the model according to the setup provided in section 4:

1: weekdays IS # AND frequency IS 1 AND Id IS #: granted with Rule Weight:1.00

2: weekdays IS 1 AND frequency 1 AND Id IS #: granted with Rule Weight:1.00

3: weekdays IS # AND frequency 2 AND Id IS 1: denied with Rule Weight: 0.98

4: weekdays IS 1 AND frequency IS 4 AND Id IS #: denied with Rule Weight: 1.00

5: weekdays IS 1 AND frequency IS 2 AND Id IS 4: denied with Rule Weight: 0.97

6: weekdays IS 4 AND frequency IS 2 AND Id IS 2: granted with Rule Weight: 0.51

7: weekdays IS 1 AND frequency IS 1 AND Id IS 1: granted with Rule Weight: 1.00

8: weekdays IS 1 AND frequency IS 1 AND Id IS 4: granted with Rule Weight: 0.99

9: weekdays IS # AND frequency IS 3 AND Id IS 3: denied with Rule Weight: 1.00

10: weekdays IS 2 AND frequency IS 1 AND Id IS 1: granted with Rule Weight: 1.00

C. Post Training Validation

In this validation phase, our training dataset is divided into 10 equal partitions where one of them (10% of the training dataset) is used as a test set and the remaining 9 (90% of the training dataset) are used as the training set. On the next round, one of the nine training partitions is picked as the test set while the test partition in the previous step is merged to the remaining 8 partitions to form a newer training set. This process is repeated 10 times until each of the partitions has been used as the test set. Finally, the results obtained in every process are averaged to obtain the global classification performance of the model over both the training and test sets.

This method is commonly known as 10-fold cross validation. By using the 10-fold cross validation, the classification performance of the model was computed by a single run of a hybridized fuzzy genetics-based machine learning algorithm. A total of 1077 records were used and the model was able to achieve classification accuracy of 99.50% and 99.16% on the training and test partitions respectively.

D. Evaluation on the Unseen Dataset (Test set)

After performing the post-training validation discussed in section D, we evaluated the model's ability to predict over the unseen data. A total of 732 records were picked from the samples to make up the dataset used at this phase. This dataset was not used during the training of the model and it is therefore an unseen, although it is from the same domain as from the training set. The classification accuracy of the model on the unseen dataset was 88.45%.

The model's Detection Rate (DR) and False Positive Rate (FPR) were computed as follows:

$$DR = \frac{TP}{TP+FN} = 99.5\%$$

where TP is the number of true positive events, actual attacker events correctly detected as an attack by the model and FN is the number of false negative events or the number *of normal events detected as an attack by the model.*

$$FPR = \frac{FP}{FP+TN} = 4.67\%$$

FP is the number of false positive or the number of normal events predicted by the model as an attack and TN is the number of true negatives, which means normal events predicted as a normal by the model.

*The model has achieved the detection rate of 99.5% and the false positive rate of 4.67%. The false positive rate of 4.67% could still be high for some applications; however, compared to linear approaches, the main advantage of this approach is that it provides highly interpretable rules which could be used to during decision making***.**

VI. CONCLUSIONS AND FUTURE WORK

Intrusion detection approaches are a less studied approach in securing RFID systems. In this paper, we made use of the hybrid fuzzy genetics-based machine learning algorithm to design an anomaly intrusion detection system for RFID systems by making use of RFID generated events in the past. The results from our experiment indicate that a high detection and low false positive rate can be achieved by the proposed model. The model also has an advantage in that it yields highly interpretable fuzzy rules that enhance human decision making when there is no clear distinction between the normal and anomalous situations.

The result from our experiments show that the hybrid fuzzy genetics-based machine learning algorithm can achieve high performance in detecting anomaly intrusions in RFID systems, specifically cloning intrusion caused by one or more cloned RFID tags in the system. Over all the detection rate of the model was 99.5% and the false positive rate was 4.67% over the training and test dataset used in the computational experiments of this paper. The results indicate that such models can successfully be adapted to other applications such as supply chain management where location and time attributes could be used to model the normal behavior of an item in the system [11].

characteristic features of the proposed fitness function are as follows:

The algorithm is capable of producing fuzzy rules which are more effective for detecting intrusion in a RFID (Table 4).

Table.4 Different algorithms performances comparison

Algorithm	False alarm rate %	Detection rate %	Complexity
FGBML	4.67	99.5	O(n)
NCP (17)	0.66	98.78	O(n)
EFRID (25)	7	98.95	O(n)
RIPPER-Artificial Anomalies (26)	2.02	94.26	O(n×log^{2}n)
SMARTSIFTER (27)	0	82	O(n^2)

One of the future extensions of this paper could be applying the method to other RFID application areas, such as supply chain management, by assessing the most relevant features to characterize the system. In this paper, we were able to detect the presence or absence of an anomaly in the system along with a given degree of confidence; however, we did not differentiate between the normal and intrusive events in the system. Extending the method so that it would differentiate between the normal and intrusive events would be another useful task to consider. In addition, assessing the applicability of this approach to identify anomalies caused

by other RFID attacks (e.g. denial of services) would be a useful endeavor for future research..

REFERENCES

[1] ABERDEEN GROUP, I., 1996-2011-last update, Aberdeen Group Research Library [June/27,2011].

[2] JIM HURLEY, J.H., May 13 , 2003, 2003-last update, Identity Theft: A $2 Trillion Criminal Industry in 2005 [June/27, 2011].

[3] WORLD HEALTH ORGANIZATION, EXECUTIVE BOARD, 18 December 2008, 2008-last update,

[4] STAAKE, T., THIESSE, F. and FLEISCH, E., 2005. Extending the EPC Network – The Potential of RFID in Anti-Counterfeiting, *ACM Symposium on Applied Computing — SAC 2005* 2005, ACM Press, pp. 1607-1612.

[5] KING, B. and ZHANG, X., 2007. Securing the Pharmaceutical Supply Chain using RFID, *Proceedings of the 2007 International Conference on Multimedia and Ubiquitous Engineering* 2007, IEEE Computer Society, pp. 23-28.

[6] WU, N.C., NYSTROM, M.A., LIN, T.R. and YU, H.C., 2006. Challenges to global RFID adoption. *Technovation,* **26**(12), pp. 1317-1323.

[7] SARMA, S., 2002. white paper: Towards the 5¢ Tag.

[8] MARY CATHERINE O'CONNOR, Aug. 23, 2010, 2010-last update, Printed-Electronics RFID Tags Debut [Homepage of RFID Journal LLC.], [Online]. Available: http://www.rfidjournal.com/article/purchase/7821 [June/2011, 2011].

[9] L. MIROWSKI, AND J. HARTNETT, 2007. Deckard: a system to detect change of RFID tag ownership. **7**, pp. 89-98.

[10] DENNING, D.E., 1987. An Intrusion-Detection Model. *Software Engineering, IEEE Transactions on,* **SE-13**(2), pp. 222-232.

[11] LEHTONEN, M., MICHAHELLES, F. and FLEISCH, E., 2007. Probabilistic Approach for Location-Based Authentication.

[12] DIMITRIOU, T., 2005. A Lightweight RFID Protocol to protect against Traceability and Cloning attacks, Security and Privacy for Emerging Areas in Communications Networks, 2005. SecureComm 2005. First International Conference on 2005, pp. 59-66.

[13] LEHTONEN, M., OSTOJIC, D., ILIC, A. and MICHAHELLES, F., 2009. Securing RFID Systems byDetecting Tag Cloning, Proceedings of the 7th International Conference on Pervasive Computing 2009, Springer-Verlag, pp. 291-308.

[14] L. MIROWSKI ET AL., 2008. A RFID Proximity Card Data Set. 2008: School of Computing and Information Systems, University of Tasmania, Hobart, Australia.

[15] ROTTER, P., 2008. A Framework for Assessing RFID System Security and Privacy Risks. Pervasive Computing, IEEE, **7**(2), pp. 70-77.

[16] JUELS, A., RIVEST, R.L. and SZYDLO, M., 2003. The blocker tag: selective blocking of RFID tags for consumer privacy, Proceedings of the 10th ACM conference on Computer and communications security 2003, ACM, pp. 103-111.

[17] ISHIBUCHI, H., NAKASHIMA, T. and KURODA, T., 1999. A hybrid fuzzy genetics-based machine learning algorithm: hybridization of Michigan approach and Pittsburgh approach, Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on 1999, pp. 296-301 vol.1.

[18] ABE, S. and MING-SHONG LAN, 1995. Fuzzy rules extraction directly from numerical data for function approximation. Systems, Man and Cybernetics, IEEE Transactions on, **25**(1), pp. 119- 129.

[19] NAUCK, D. and KRUSE, R., 1999. Obtaining interpretable fuzzy classification rules from medical data. Artificial Intelligence in Medicine, **16**(2), pp. 149-169.

[20] NOZAKI, K., ISHIBUCHI, H. and TANAKA, H., 1996. Adaptive fuzzy rule-based classification systems. Fuzzy Systems, IEEE Transactions on, **4**(3), pp. 238-250.

[21] ISHIBUCHI, H., NOZAKI, K. and TANAKA, H., 1992. Pattern classification by distributed representation of fuzzy rules, Fuzzy Systems, 1992., IEEE International Conference on 1992, pp. 643-650.

[22] ISHIBUCHI, H., 2007. Evolutionary Multiobjective Design of Fuzzy Rule-Based Systems, Foundations of Computational Intelligence, 2007. FOCI 2007. IEEE Symposium on 2007, pp. 9-16.

[23] ABADEH, M.S., HABIBI, J. and LUCAS, C., 2007. Intrusion detection using a fuzzy genetics-based learning algorithm. J.Netw.Comput.Appl., **30**(1), pp. 414-428.

[24] FULCHER, J. and JAIN, L.C., 2008. Computational Intelligence: A Compendium. 1st edn. Springer Publishing Company, Incorporated

Aspects regarding the relevant components of online and blended courses

A. Naaji, A. Mustea, C. Holotescu and C. Herman

Abstract— In the last years a diversification of teaching methods has been observed, mainly determined by the growing ubiquity of Social Media, the emerging mobile technologies and the augmented reality. Major challenges in education that involve tremendous development and innovation are blended courses/ flipped classrooms integrating Social Media (SM), Open Educational Resources (OER) and Massive Open Online Courses (MOOC) [1].

This paper is focused on evaluating the e-learning experiences of various actors in Romanian educational system. There is a tendency to use virtual learning environments with increasing frequency in higher education, many participants experiencing both online and blended courses. Another issue approached in the paper concerns the relevance of components of online/ blended courses. In this context, the paper analyzes the importance of these elements in relation to various fields, such as: exact sciences, social sciences, humanistic studies, medical sciences, etc.

In conclusion, we identify the most relevant elements in the development of an online/ blended course for various domains. The results will emphasize the standards required in evaluating the quality of online and blended courses.

Keywords—e-learning, blended learning, MOOCs, online course components, e-learning experience

I. Introduction

THE development of Internet and the new approaches to higher education in terms of teaching methods led to a similar development of online education, also known as e-learning. In Romanian educational system only a blended learning approach is accepted by agency responsible with the accreditation of educational programs. This is an educational form mediating the transition towards exclusively online courses. Nevertheless the participants in the educational system had the opportunity to participate in online courses in the form of MOOCs or in training programs offered for professional development.

The paradigm shift from traditional face-to-face education to e-learning comes with several challenges. One of the main challenges concerns the assessment of the quality of e-learning [2], [3], [4]. This subject is still debatable worldwide. There is an intense study regarding several quality models proposed by various institutions. Therefore it is important for the Romanian cultural context to identify and propose a quality standard to be used in higher education. In order to develop adequate standards, the experience of online course participants needs to be considered.

In this context course templates need to be adapted to online education. In order to achieve the quality standards several course elements will be considered. On the other hand, the learner-teacher interactions are distinct from traditional education ones, as are the roles and involvement assumed by teachers and learners. The structure and content organization necessary for an optimal e-learning experience and efficiency also have specific requirements, both pedagogical and technical.

With all these challenges, the paradigm shift from traditional to online education can be confusing for both teachers and learners. Most comparisons between e-learning and traditional education have emphasized that it is not the environment, but the instructional design and quality of the online instruction that ensures the course efficiency and students' learning [5].

A. E-learning forms and characteristics

There are numerous definitions of e-learning [6], therefore in order to avoid confusion we will define e-learning in accordance with the purpose of this paper. Agreeing with R.C. Clark and R.E. Mayer, "we define e-learning as instruction delivered on a digital device such as a computer or mobile device that is intended to support learning." [5].

In defining the quality of e-learning, three main components were identified: technology, learning content and learning design [6]. Combining technology, learning content and learning design will result various taxonomies or forms of e-learning. For example, K. Fee identifies five models of e-learning: online courses, integrated online and offline learning, self-managed e-learning, live e-learning and electronic performance support [6]. In this paper, we will consider all three forms of online education familiar to our participants: online courses, blended learning (combining both online

A. Naaji is with Department of Computer Science, "Vasile Goldis" Western University, 310025 Arad, Romania (+40-257-285813; e-mail: anaaji@uvvg.ro).

A. Mustea is with Department of Psycho-pedagogy, "Vasile Goldis" Western University, 310025 Arad, Romania (e-mail: anca.c.mustea@gmail.com).

C. Holotescu is with Department of Computer Science, University Politehnica Timisoara, 300006 Timisoara, Romania. (e-mail: carmen.holotescu@upt.ro).

C. Herman is with Department of Information Technology, "Vasile Goldis" Western University, 310025 Arad, Romania (e-mail: cosmin@uvvg.ro).

activities and face-to-face traditional methods) and courses designed for MOOC platforms [2], [7].

The most challenging aspect in developing courses for all forms of e-learning is the instructional design. Here is where confusion mostly occurs, since e-learning overcomes the simple transposition of traditional courses in digital form. It is also important to consider the differences between pedagogy, adult learning and online learning, as each involves a different educational design [8]. E-learning (*allagegogy* meaning "teaching to transform") focuses on the learner's independence and the changing nature of lifelong learning [8].

The characteristics of an online learner are essential for adequately choosing course content and developing online courses by considering the characteristics of e-learning instructional design. The online learner has the following characteristics: ability to work independently and in groups, responsibility in completing assignments and readings, ability to learn using content in various formats, time management and personal organization skills, and the knowledge and skills to use technology [8].

Other important aspects for developing online courses are: the structure and the components of a course, the multimedia resources, the teacher-learner and the learner-learner interaction, the presentation/ delivery mode, and the role and selection of assessment methodologies [9].

B. Quality of e-learning

There are numerous aspects to consider when developing an online/ blended course, some of which we mentioned in the previous section. Going beyond the characteristics of e-learning, there is the issue of quality in online education. Several models of what is e-learning quality were developed over the last decade [10], [11].

For example, an e-learning quality guide in higher education developed by the Swedish National Agency for Higher Education [11] proposes ten crucial aspects to consider when evaluating quality in e-learning: (1) material/ content, (2) structure/ virtual environment, (3) communication, cooperation and interactivity, (4) student assessment, (5) flexibility and adaptability, (6) support (student and staff), (7) staff qualification and experience, (8) vision and institutional leadership, (9) resource allocation and, (10) the holistic and process aspect. Other models might emphasize different aspects, such as technology or examination security [10], but there are also common important elements such as institutional vision, instructional design, course structure, student and staff support, and student assessment.

There are several guidelines of what constitutes quality in online learning, being online courses, blended learning or MOOCs. But specific guidelines for the Romanian educational system need to be refined, considering both common elements of international guides concerning e-learning quality, and the cultural context with its constraints. This requirement is becoming more and more stringent as the development of online learning gets exponential.

C. The development of online learning

Worldwide there is a permanent and rapid development of online learning, evidenced by various statistics both in academic and entrepreneurial settings. R. Davis and D. Wong noted that if in 2001 approximately 3 million people were involved in some form of e-learning, in 2006 their number increased to 6 million people [3]. Statistics concerning an American north-eastern university emphasize a growth from 4 online courses in 1996 to over 500 courses in 2010 [2].

An increase in using e-learning for training was also noted in the work force. R.C. Clark and R.E. Mayer [5] present the situation of training delivery via computer in the United States: approximately 11% in 2001, 29% in 2006 and 36.5% at the beginning of 2010.

By 2011 the massive open online courses movement was materialized in the form of multiple platforms offering open courses to large numbers of students: Coursera, edX, Udacity [12], [13]. The 2011 fall AI online class at Standford University registered 160,000 students, 23,000 students completing the ten-week course [12]. Coursera posted more than 200 free courses taught by professors from more than 30 top world universities [12]. Currently Coursera has 538 courses posted in English only. European MOOC platforms, such as Iversity, are also developing [7].

The situation of e-learning in Romania follows a similar trend, even though on a smaller scale. Most Romanian universities use proprietary or open-source LMSs for online/blended courses or for courses enhancement. During the last years a high percentage of the pre-university and academic teachers took part in online/ blended trainings related to eLearning offered in projects co-financed with European funds. Also Romania appears active in the OER movement, mainly through initiatives by institutions/groups and engaged individuals and through specific projects or programmes, but also there are a few initiatives at the government level that can become driving forces [14].

While most studies focus on the assessment of quality and efficiency of e-learning, few of them consider the experience of the actors involved. Therefore, we believe it is important to comparatively see how e-learning, blended learning and traditional education are perceived by learners and teachers.

II. METHODOLOGY

A. Objectives and hypothesis

The purpose of this study is mostly an exploratory one, more precisely to investigate the importance of different elements used in building and facilitating blended courses from the perspective of all types of actors involved in the Romanian educational settings. A previous research [9] illustrated the importance of online course elements for students, for which further investigation that includes teachers, administrative staff and student was required.

The first aspect explored in this study was the e-learning experience of participants. Further, we tested two hypotheses: (1) There are course elements significantly more important for all participants; and (2) The importance of course elements differs across domains.

B. Participants

There were 84 participants in this study, the majority of them working as teachers, researchers or administrative staff in undergraduate or graduate educational institutions, with 41.7% males and 58.3% females. Seven participants are high school or undergraduate students.

Most participants were graduates (39.3%) or postgraduates (54.8%), and just five of them (6%) were high school graduates only. Forty five participants were related to hard sciences (exact sciences) (53.6%), while other scientific affiliations were as follows: 7.1% social sciences, 9.5% economical sciences, 11.9% humanistic studies, 1.2% medical sciences, and 16.7% other domains.

C. Instruments

An online questionnaire was developed in order to gather the demographical data and assess the importance of course elements and e-learning experience. Nineteen course elements were assessed using a Likert scale with five points, with 1 meaning "not at all important" and 5 meaning "very important".

D. Procedure

The participants were invited to participate in the study via a Google form. The form was designed and developed using the standard features. The public link of the questionnaire was shared via academic networks of the authors' universities, relevant academic mailing lists, personal learning networks and social web platforms. The recommended time to complete the form was between 5-10 minutes. The questionnaire was completed online by all participants, ensuring their anonymity and confidentiality.

III. Results

An important aspect of data analysis concerns the e-learning experience of our participants. Therefore, the first part of data analysis consisted of comparing the different types of experiences. In Figure 1 the percentages of participants checking each type of answer are illustrated. Thus, we can say that only 11.9 percent had no experiences involving e-learning. From those who had some kind of e-learning experience, most participated in online courses (65.5%), blended learning (34.5%) or Massive Open Online Courses (32.1%); the percentage of those participating in MOOCs is quite impressive for the interest in this trending model for personal and professional development. It is worthy to note that almost one third of responders have experience in facilitating online courses (31%) and blended courses (28.6%), demonstrating the increasing rate of e-learning integration in Romanian education. Nevertheless, we can observe that online courses represented the most common e-learning experience. It is important to mention that all percentages are relative to the total number of participants.

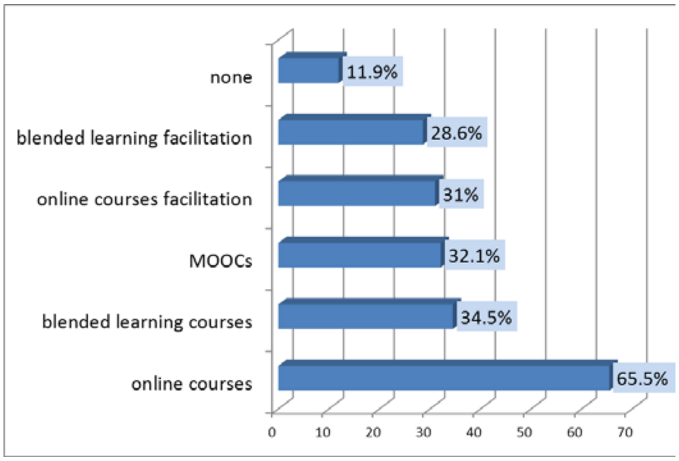


Fig. 1 e-learning experience

The importance of course elements was assessed on a five-point Likert scale, with 1 meaning "not at all important" and 5 meaning "very important". Nineteen elements of an online course were evaluated on this scale. The results synthesized in Figure 2 illustrate the importance of each element, for all participants in this study, regardless of their e-learning experience.

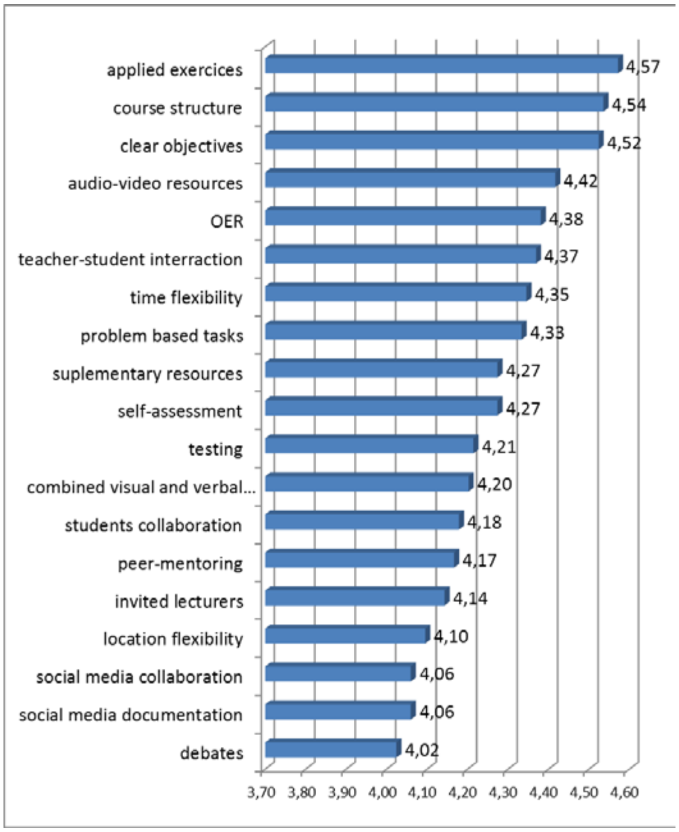


Fig. 2 the mean results, representing the importance of online/blended course components, for all participants

Thus, the most important elements of a course were considered to be: applied exercises (M = 4.57), course structure (M = 4.54) and clear objectives (M = 4.52). The least important elements of an online course were debates (M = 4.02), social media documentation (M = 4.06) and social media collaboration (M = 4.06). Nevertheless, we can observe that, on average, all the elements of an online/ blended course were considered as being important – all means represented in Figure 2 are above four points on the Likert scale.

Analyzing the significance of these differences, using the Analysis of Variance, we obtained, F = 5.508 at a significance level lower than .001. Therefore, there are significant differences between the importance attributed to the various elements of an online/ blended course.

Furthermore, using Bonferroni post hoc comparisons, we analyzed which are the elements of an online course considered significantly more important and those significantly less important. Applied exercises are significantly more important than debates (p = .000), students collaboration (p = .006), location flexibility (p = .006), testing (p = .023), social media documentation (p = .001), social media collaboration (p = .001), invited lecturers (p = .001), and peer mentoring (p = .001). The course structure is significantly more important than debates (p = .002), students collaboration (p = .017), social media documentation (p = .000), social media collaboration (p = .001), invited lecturers (p = .002), and peer mentoring (p = .030). Clear objectives are also significantly more important than debates (p = .013), students collaboration (p = .044), social media documentation (p = .001), social media collaboration (p = .004), invited lecturers (p = .016). There are no statistically significant differences between audio-video resources and any other elements of an online course. There are significant differences only between the elements situated at the two ends of the importance scale. Summarizing these results we can say that all components evaluated are considered important and only few of them were considered significantly different in importance.

Another comparison we made was related to the importance of course elements for the various fields of expertise: exact sciences, social sciences, economic sciences, humanistic studies, medical sciences, and other domains. Since there was only one participant for medical sciences, we excluded his results from this comparison, in order to be able to analyze the statistical significance of observed differences.

The average importance of course elements for each domain is illustrated in Figure 3. The results emphasize the differences in the importance attributed to various course elements. Thus, the elements considered more important by the participants from exact sciences are: peer-mentoring, invited lecturers, social media collaboration, social media documentation, open educational resources, testing, problem-based tasks, students' collaboration, self-assessment, debates, course structure and clear objectives. The participants from social sciences considered audio-video resources and teacher-student interaction as being more important than they were for the participants from all other domains. The elements more important for participants from economic sciences were applied exercises, time flexibility and location flexibility. Supplementary resources were equally important for the participants from exact sciences and economic sciences.

When testing the statistical significance for these differences using the Kruskal-Wallis test, only two of the course elements proved to be significantly different in importance as a function of the activity domain: clear objectives (chi square = 9.260, p = .026) and problem based tasks (chi square = 11.279, p = .010). A significance level between .1 and .05 was also found for self-assessment, time flexibility, location flexibility, testing, social media documentation, social media collaboration, invited lecturers. Further research on more homogenous groups might find these elements as also being significantly different across domains.

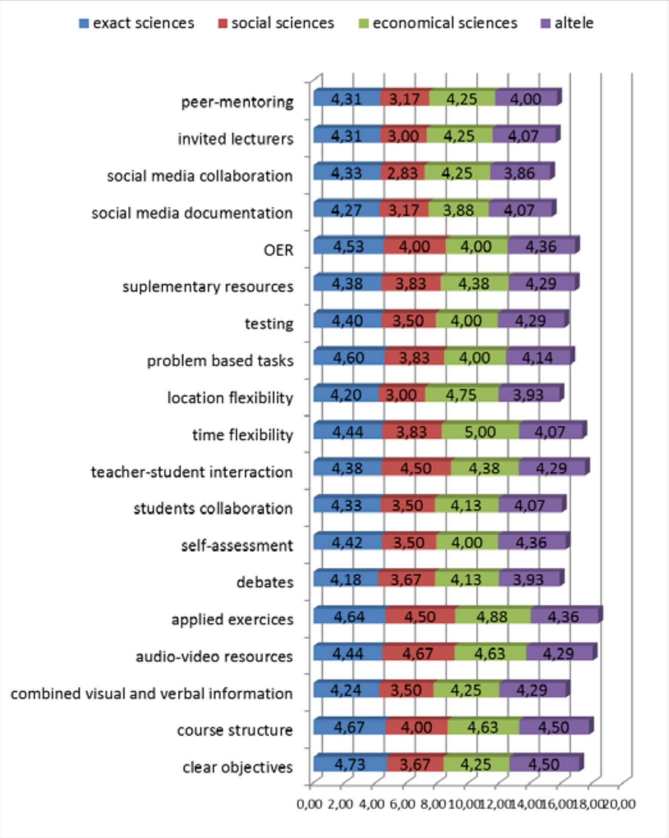


Fig. 3 e-learning components assessment by domain

The questionnaire has also two open-ended questions asking respondents to list/ identify main aspects/ advantages and constraints to uptake in an online/bleded course, almost all participants sharing their impressions.

The main advantages expressed by participants are listed below:

- possibility to integrate new technologies and pedagogies;
- mobile learning, flexibility in time and space, the possibility to learn and to interact with materials and peers anytime and everywhere, using mobile devices, in own pace;

- real-time feedback from teachers, peers, external learners and practitioners;
- a large diversity of resources, adapted to all learning styles;
- access to quality OER and to updated/multimedia resources posted on social networks;
- continuous self-assessment;
- possibility to focus on new topics using the previous/ tacit knowledge;
- student centered learning;
- interactivity, imagination, critical thinking and group work are stimulated;
- digital skills are improved;
- the possibility to build your Personal Learning Environments;
- learning communities are nurtured and can continue after the end of course.

The disadvantages of online/blended courses noted by participants are summarized in the following paragraph:

- a national policy related to the integration of new technologies/ pedagogies, OER in education doesn't exist;
- teachers should be trained to be able to develop and facilitate online and blended courses;
- there are no incentives to reward teachers using open technologies and pedagogies;
- student assessment when using online collaboration and social media could be difficult;
- there should be a team of experts to develop a quality online courses;
- the missing of digital skills of both students and teachers could be a barrier for such courses;
- the management of time could be a challenge;
- the lack of feedback from teachers could demotivate students.

IV. Conclusions

Considering the rapid development of e-learning worldwide as well as in Romanian context, there is necessary to identify the main and quintessential elements of an online/blended course and what defines the e-learning quality.

In this research we aimed to identify the importance of the components of an online/blended course. We started with the idea that some of the proposed elements will be considered more important, and there will be differences across domains.

Participants varied in what concerns the e-learning experience, educational background as well as area of specialization. For all participants, the elements of an online/ blended course proposed in the questionnaire were considered important, all means being above 4 on a five points Likert scale. Nevertheless, three course elements were considered significantly more important: applied exercises, course structure, and clear objective. Differences across domains were also identified (clear objectives and problem based tasks).

These results emphasize and define the directions of future research concerning the e-learning quality and the experiences of all actors involved in e-learning in Romanian educational system. The need for clear policy regarding the integration of new technology in education is revealed by this study: both by quantitative and qualitative results. Further studies are needed in order to develop clear and comprehensive guidelines for qualitative online/blended courses. Another important need resides in supporting and informing both, students and staff, about the quality standards of online education.

References

[1] L. Johnson, S. Adams Becker, V. Estrada, A. Freeman, *NMC Horizon Report: 2014 Higher Education Edition*, Austin, Texas: The New Media Consortium, 2014, Retrieved from http://www.nmc.org/pdf/2014-nmc-horizon-report-he-EN.pdf.

[2] D. Lewis and E. Chen, "Factors Leading to a Quality E-Learning Experience" in *T. Kidd, Online Education and Adult Learning: New Frontiers for Teaching Practice*, New York, USA: Information Science Reference, 2010.

[3] R. Davis and D. Wong, "Conceptualizing and Measuring the Optimal Experience of the eLearning Environment", in *Decision Sciences Journal of Innovative Education*, vol. 5, pp. 97-126, Jan. 2007.

[4] R. Donnelly and K.C. O'Rourke, "What now? Evaluating eLearning CPD practice in Irish third-level education", in *Journal of Further and Higher Education*, vol. 31, pp. 31-40, Feb. 2007.

[5] R.C. Clark and R.E. Mayer, *e-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning*, 3rd ed., San Francisco, USA: John Wiley & Sons, 2011.

[6] K. Fee, *Delivering E-Learning: A complete strategy for design, application and assessment*, London, United Kingdom: Kogan Page, 2009, p.14.

[7] A. Mustea, A. Naaji. And C. Herman, "Using Moodle for the development of Massive Open Online Courses", presented at the 10th International Scientific Conference eLearning and software for Education, Bucharest, 2014, 10.12753/2066-026X-14-000

[8] L.A. Tomei, "A Theoretical Model for Designing Online Education in Support of Lifelong Learning", in T. Kidd, *Online Education and Adult Learning: New Frontiers for Teaching Practice*, New York, USA: Information Science Reference, 2010.

[9] A. Naaji, A. Mustea, C. Herman, "Implementation Model for New Technologies in Online Education", in *Proceedings of the 24th EAEEIE Annual Conference*, Crete, Greece, 2013, ISBN: 978-1-4799-0042-8, p.76-80.

[10] S. Uvalic-trumbic, Sir J. Daniel, *A Guide to Quality in Online Learning*, Academic Partnership, 2013.

[11] Swedish National Agency for Higher Education, *E-learning quality: Aspects and criteria for evaluation of e-learning in higher education*, Stockholm, Sweden: Högskoleverkets, 2008.

[12] F.G. Martin, "Will Massive Open Online Courses Change How We Teach?", in *Communications of the ACM*, Vol. 55, no. 8, 2012.

[13] J. Leber, "The Technology of Massive Open Online Courses", MIT Technology Review, vol. 116, no. 1, 2012.

[14] C. Holotescu, *OER in Romania. Report in POERUP Project: Policies for OER Uptake*, 2012, retrieved from http://poerup.referata.com/wiki/Romania.

Improvement of QoS in Grid Computing by Combination Heuristic Algorithms

E. Tavakol, M. Fathi,S. Navaezadeh

***Abstract*—** Grid computing is a new technology, which follows the goal of distributing resources and cooperating in large scale. Task scheduling in order to reach to desired quality is one of the important fields in grid environment. Although many efforts have been done about scheduling ways and distributing task in grid area, grid applications need guarantee in order to provide the grid quality service, which can be arrived by reservation. Reservation includes advance reservation and immediate reservation. Reservation algorithm must be executed in a way that minimizes the delay of accepted requests and maximize the efficiency of resource usage (efficiency of resources).

The suggested algorithm in this paper applies the combination of Simulated Annealing algorithm (SA) and particle swarm optimization algorithm (PSO). In addition, of doing the task in the least time, it provides the load balance for the resources in the grid. Therefore by using this algorithm, two supper important QoS factors (efficiency and load balance) are reached as it's possible and task scheduling and resource allocation is done by regarding these two factors..

***Keywords*—**load balance ,Quality of service, Executing time, Reservation, PSO, SA.

I. INTRODUCTION

During the recent years, some methods based on the collective flying style of the birds are regarded as a solution for combinative optimization. In fact, a new method in computing by the bird’s action that is called the particle swarm optimization is invented. The main features are positive feedback and distribute computation. Grid computing is a software and hardware ultra structure, which bring wide, stable accessibility to the resources in the network [1]. Basic goal in grid computing is to use the processing power of existing computers of the network to solve the complicated time consuming problems. In order to reach to the goal of grid and maximum usage of the existing resources in area of grid, the way the tasks are distributed between resources and their

Scheduling has a very important place. This scheduling is done by regarding the quality of service and it always tries to share the tasks between resources in away to maximize QoS results [2, 3, 4].

The number of parameters is too much. We can mention Reliability, Availability, and Throughput [5]. Many algorithms related to task scheduling in distributed system and in grid are presented[6,7,8,9,10,11,12,13].

The goal of these algorithms is to optimize a factor of QoS but using of different resources in grid systems had its own problems, which appeared by passing of the time and finding solution for these problems led grid to develop and complete.

One of these problems is that the resources are in different management areas, which has its local scheduler. It is possible that while the assigned task is executing in that resource in grid system, the local scheduler ends the task over the resource and the assigned task be left unfinished by grid system.

Advanced reservation and immediate reservation are the effective technologies to provide quality of service in grid.

An advanced reservation is a limited procurator of abilities of certain resource during a certain time, which is given to applicant through a negotiation process from resource owner.

Immediate reservation can be considered as an advanced reservation which starting time is just now. By increasing workload or number of requests, in certain times we many see using of one of resources while in other time the resources are idle.

Requests in grid environment are usually flexible to their start and ending time and proper usage of request flexibility from reservation algorithm leads increasing resources optimization and reduces the number of rejected requests. Advances reservation allows the users to use enough resource to execute application at the same time in order to reserve needed resource in an advanced reservation system.

Elham Tavakol received the B.Sc. degree in Computer Engineering from Islamic Azad University, Mahshahr Branch in 2006, and the M.Sc. degree in Computer Engineering from Science and Research Khouzestan Branch in 2008 and 2011, respectively. here research interests include Grid Computting and PSO. Sama Technical and Vocational Training College, Islamic Azad University, Mahshahr, Branch Mahshahr, Iran, email: elham.tavakol90@gmail.com

Mahsa Fathi received the B.Sc. degree in Computer Engineering from Islamic Azad University, Mahshahr Branch in 2006, and the M.Sc. degree in Computer Engineering from Science and Research Khouzestan Branch in 2008 and 2011, respectively. here research interests include Grid Computing and fuzzy logic. Department of computer engineering, Ahvaz branch, Islamic Azad University, Ahvaz, Iran, email: mahsa_fathi2013@yahoo.com

Sedigheh Navaezadeh received the B.Sc. degree in Computer Engineering from Islamic Azad University, Mahshahr Branch in 2008, and the M.Sc. degree in Computer Engineering from Science and Research Khouzestan Branch in 2010 and 2013, respectively. Here research interests include Grid Computing . Sama Technical and Vocational Training College, Islamic Azad University, Mahshahr, Branch Mahshahr, Iran, email: Snavaezade92@yahoo.com

User must define its request by some parameters and give it to system [14] then advanced reservation system inspects its possibility of acceptance. Evaluating acceptance possibility of this request, regarding flexibility parameters in different algorithms happens in different ways. Regarding the point that the request scheduling is complete-NP [15].it is impossible to use an algorithm that evaluates the whole solutions and choose the best solution. Thus, we have to use awarding and heuristics algorithm. The optimizing algorithm of particle swarm is of one of the heuristics algorithms that are modeled from bird collective movement. This algorithm can reach to an optimized solution in the problem solving environment. In this paper, the SA algorithm is used to optimize algorithm of particle Swarm as coefficient so that we speed up the convergent. Because in advanced reservation its needed to answer the requests in limited time. Thus, combination of these two algorithms makes the advanced reservation extensible and can find an optimal answer, not necessarily optimum solution for the problem in high workload and high number of resources.

Our motivation from this research is to combine two SA and PSO algorithm that outcome result shows that QoS parameters by using this combination are improved remarkably.

The remainder of this paper has the following structure: In Section 2 related work of advance reservation and grid QoS is reviewed; In Section 3 combining of PSO and SA for Advance reservation is proposed; In Section 4, results of experiment are presented and discussed; In Section 5 we make a conclusion and look forward to futurework.

II. RKROW DETALE

In order to define time of processor for immediate soft tasks the mechanism, which provides the request status of advances reservation to run in a certain starting time, will be so functional and important.

Too many researchers have been done about back up field from advanced reservation about network quality service and grid resources. In case of special network application ,video conference, presenting from far distance, multiuser games and etc that users tend to have an advanced reservation and provide themselves a level of quality service Ber**os**n and friends [16] present an architectural based on server which on that, unless the reservation isn't activated. Multi section directing or any special reservation isn't needed. This architectural based in server also allows applicants to request for advanced reservation, without being activated during advanced reservation time. Foster and friends [17] presented Globus architectural for specializing and reservation. This architectural tries to research on service quality of and to and in newly appeared applicants based in network. Gara did these matter (case) using dynamic discovery then immediate and advanced reservation resources that normally aren't the same and controlled and managed independently. Buyyaa [18] presents a structure based on array for managing of reservation in grid computing. Idea of this method is specifically inspired from line and tree calendar.

Regarding the task record, we get the concept of advanced reservation, fairly in field processor schedulin**g.**

III. PARTICLE SWARM OPTIMIZATION AND SIMULATED ANNEALING

I. Basic PSO Algorithm

The concept of PSO roots from the social behavior of organisms such as bird flocking and fishing schooling. Through cooperation between individuals, the group often can achieve their goal efficiently and effectively. PSO simulates this social behavior as an optimization tool to solve some optimization problems. In a PSO system, each particle having two properties of position and velocity represents a candidate solution to the expressed by the objective function. In the iteration, the objective function is calculated to establish the fitness value of each particle using position as input. Fitness value determines which position is better. Each particle flies in the search space with a velocity that is dynamically adjusted based on its own flying experience and its companions' flying experience. In other word, every particle will utilize both the present best Position information of its own (Pbest) and the global best position information (gbest) that the swarm has searched up-to now to change its velocity and thus arrives in the new position.

PSO can be described mathematically as follows.

Suppose that the search space is of d-dimension and the number of particles is n. The ith particle is represented by a dimension

Vector Xi=(xi1,xi2,…..,xi$_{\mathbf{d}}$)

Pbesti = (p1,p2,….p$_{\mathbf{D}}$) denotes the best position searched by the ith particle and the gbest = (g1,g2,……,gn) is the best position Searched by the whole swarm up-to-now. Each particle updates its velocity and position according the following equations.

Vid = w × Vid +c1 × rand () (**P**bestid – Xid) + c2 × rand () (gbestid –Xid) ………………………………………(1)

Xid = Xid + Vid ....(2)

Where w is the inertia coefficient which is a chosen constant in interval [0,1]; c1,c2 are two acceleration constants; rand() is random value in interval [0,1]. The velocities of particles are restricted in interval [Vmin, Vmax]. If the resulting value is smaller than Vmin, one element of the velocity vector is set to

Vmin; if the resulting value is greater than Vmax, one element of velocity vector is set to Vmax.

II. Simulated Annealing Algorithm

The term annealing derives from the physical process of heating and then slowly cooling a substance to obtain a strong crystalline structure. Simulated Annealing (SA) is the optimization method to stochastically simulate this physical process of annealing on the computers. In SA, the simulation proceeds by randomly generating a solution and then determining its acceptance in certain probability. A temperature parameter is used to determine this probability. In the basic algorithm of SA, three operations bear an important role: generate, accept and cool.

The general operation modifies a current solution X and generates a next solution X' using a probability distribution G(X,X').

The accept operation is a judgment to decide whether to accept the modification or not. The acceptance of modification is determined from a difference ΔE (=E'-E) of a current energy E=f(x) and a modified energy E'=f(X'), and a temperature parameter T. If the ΔE≤ 0, the modification is accepted. In case ΔE > 0, the modification is accepted at certain probability. This algorithm can be defined as:

$$\text{Paccept}=\begin{cases}1 & \text{if} \quad \Delta E\leq 0\\ \text{Exp}\ (-\Delta E\ /\ T\ ) & \text{otherwise}\end{cases} \quad \ldots\ldots\ldots.(3)$$

The temperature T is the important parameter to control the acceptance of the modified solution. At the beginning of the simulation, both the temperature and the acceptance levels are high. As the simulation proceeds and the temperature decreases, solutions that have the bigger fitness values.

Cooling is the operation to generate the temperature of the next state Tk+1 from the temperature of the current state Tk.

The simulation begins with the initial solution x with energy of E and the initial temperature T. The solution is then randomly modified to x' with energy of E'. The acceptance of the modification is calculated from the difference of energy **,** ΔE(= E' − E), and the temperature Tk. If accepted, the solution x' becomes the starting point of the next step. These operations are repeated long enough at each temperature for the system to

reach a steady state, or equilibrium. When reaching the steady state at Tk**,** the temperature is cooled to Tk+1 and the simulation

is repeated until reaching steady state again. Users define the terminal criterion, such as the number of evaluations. The simulation is concluded when the temperature becomes low enough and a terminal criterion is met.The temperature can be updated by the following method

$$Tk = \alpha\ .\ Tk\text{-}1 \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(4)$$

The problem involves dividing the circuit net list into two subsets. The objective function captures the interconnection information. The mathematical representation of the objective function is given as

$$F=\sum_{i=0}^{k} [\![1/m]\!] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..\ldots\ldots.(5)$$

F= Fitness Function

M= Min Cut.

III. OSP_SA Hybrid Algorithm

This section describes the proposed hybrid algorithm consisting of particle swarm optimization (PSO) and simulated annealing (SA) algorithms. This problem is very complex in nature and difficult to solve large-scale problems by optimization techniques. In this case, it is well experienced that meta-heuristic algorithms can often outperform conventional optimization methods when applied to difficult real-world problems. An encoding scheme is first presented in order to generate schedules. Then SA and PSO are separately reviewed and designed. Ultimately, the hybridization procedure of these algorithms is explained.

The idea of the proposed hybrid algorithm, called PSO-**AS**based on PSO and SA algorithms for scheduling problems, has been widely exploited in the literature [6, 7, 34, 35].PSO possesses high search efficiency by combining local search (by self experience) and global search (by neighboring experience). Moreover, SA is meta-heuristic, that is, designed for finding a near optimal solution of combinatorial optimization problems. Therefore, the PSO and SA algorithms are combined which can omit the concrete velocity–displacement updating method in the traditional PSO for scheduling problem.

The proposed hybrid algorithm includes two phases: (1(the initial solutions are randomly generated and (2) the PSO algorithm combined with the SA algorithm is run. The general outline of the hybrid algorithm is summarized as follows:

```
(1) iter ←0, cpt ← 0, Initialize swarm size particles
(2) stop criterion maximum number of function evaluations or
Optimal solution is not attained
(3) while Not stop criterion do
(4)for each particle i ← 1 to swarm size do
(5) evaluate (particle(i)) if the fitness value is better than the best fitness value (cbest) in history then
(6) Update current value as the new cbest.
(7) end
(8) end
(9) Choose the particle with the best fitness value in the neighborhood (gbest)
(10)for each particle i ← 1 to swarm size do
(11)Update particle velocity according to Equation (1)
(12)Enforce velocity bounds
(13)Update particle position according to Equation (2)
(14)Enforce particle bounds
(15)end
(16)if there is no improvement of global best solution then
(17)cpt ← cpt + 1
(18)end
(19)Update global best solution
(20)cpt ← 0
(21)if cpt = K then
(22)cpt ← 0
```

```
    //Apply SA to global best solution
    iterSA ← 0, Initialize T
    current solution ← global best solution
    current cost ←evaluate(current solution)
    while Not SA stop criterion do
        while inner-loop stop criterion do
            Neighbor ←generate(current solution)
            Neighbor cost ← evaluate(Neighbor)
            if Accept(current cost, Neighbor cost, T) then
                current solution ← Neighbor
                current cost ← Neighbor cost
            end
            iterSA ← iterSA + 1
            Update (global best solution)
        end
        Update(T) according to Equation(3)
        Update (SA stop criterion)
    end
end
iter ← iter + 1, Update (stop criterion)
End.
```

After selecting (proper) allocate, in order to minimize number of rejected requests, another algorithm is pre-stented that tries to offer an allocation with amount of de-lay in last declared ending time are a (td) to request the rejected requests and call this algorithm the replacement algorithm and by using it try to maximize the utility of new allocation.

In replacement algorithm we try to avoid their rejection by replacing and shifting in time dimension of rejected requests in declared boundary by requests it selves (dt - st), prevent their rejection and maximize mass of accepted requests and minimize the delayed mass,

In replacement algorithm all of rejected requests in x level will be placed and listed in order of starting time (st), then we will do the mixing on them for several times. If we have N rejected requests, it will have n! Choice.

Will have by mixing this N request for k times (k could be linear function of N) trying to linear accruing time number of rejected requests.

IV. EXPERIMENT All RESULTS

To evaluate the result, in this part an experiment is been set and the result of PSO-SA algorithm is compared with PSO algorithm and ordering algorithm. In this article it's been tried to have the comparison based on one of the QOS parameters like possessed resource mass and this parameter for each 3 cases: length, average requested resource number and ability of delay tolerance is regarded.

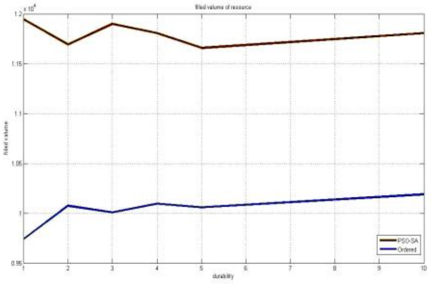


Chart 1: possessed mass by flexibility

As it's observed in chart1, the possessed mass algorithm leads that when requests flexibility increases, algorithms by sacrificing number for mass, try to maximize rate of resource use. Totally it could be seen that the resource use rate change around 17 to 25 percent for different flexibility.

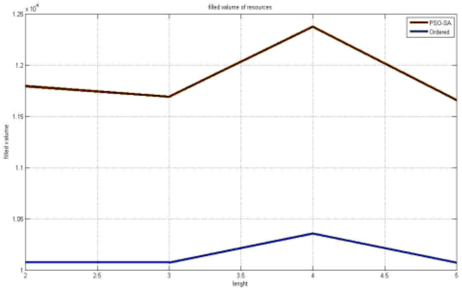


Chart2: filled mass of resources by length request average

It could be taken from result of chart2 in average length4, algorithm uses the maximum capacity delay of requests and maximum mass so It will be filled with fairly big requests and this causes the increase of possessed mass and reduces acceptance number.

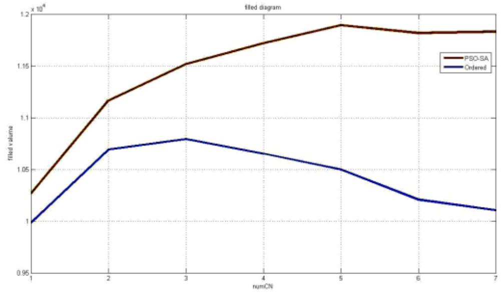


Chart 3: useful possessed mass percentage of resources by number of requested resources of requests.

In chart 3 it's observed as far as the average number of requested resource increases, possessed mass difference increases. Because in upper requested number of resource an incorrect allocation has far move bad effect to lesser number of requesting resource .as we can see difference efficiency of 2 charts in upper resource number is shown.

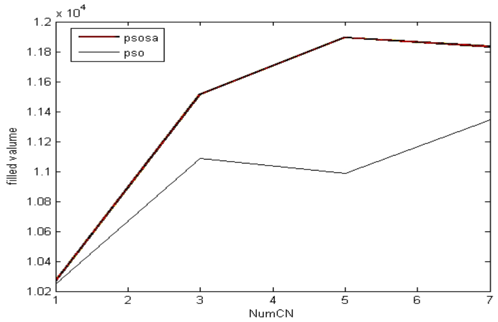


Chart4: comparison of possessed mass in 2 algorithms.

In chart4, it is observed the accepted mass is about 4% difference that shows the advantages of combination these two algorithms.

IV. CONCLUSION

In this article, we have used the combination of PSO-SA algorithm so that the resource advanced reservation problem. In heavy loads, that some requests have to be rejected must be done in optimized way. Being optimized is regarded as maximum usage of resources.

In addition, a comparison between PSO reservation and PSO-SA is done.

The result of stimulating shows the combination of these two algorithms improved PSO algorithm (even though little) this matter regarding to the little bit computing load that SA algorithm has to system is very useful.

In this paper we consider resource the same but it could done for different resource. Also it could add to priority and the importance plus the possessed mass of resources, be considered the requests priority. Also that could be two phased , it means when the work load is low, usual algorithm be use advanced, heuristic algorithm and by this act the top load of process will be less.

REFERENCES

[1]]Foster I., Kesselman C., The Grid: Blueprint for a New computing Infrastructure, Morgan Kauffman, 1999.

[2] Plestys R., Vilutis G., Sandonavicius D., "the Measurement of Grid QoS Parameters", Proceedings of the ITI 2007; 29th Int. Conf. on Information Technology Interfaces, Caveat, Croatia, June 2007.

[3] Sun X-He., Wu M., "Quality of Service of Grid Computing: Resource Sharing", The Sixth International Conference on Grid and Cooperative Computing (GCC 2007).

[4] Wang X., Luo J., Architecture of Grid Resource Allocation Management Based on QoS, Book chapter of “Grid and Cooperative Computing ,Springer Berlin Heidelberg, pp. 81-88, 2004.

[5] Plestys, R., Vilutis, G., Sandonavicius, D., "The Measurement of Grid QoS Parameters"; Proceedings of the ITI 2007; 29th Int. Conf. on Information Technology Interfaces, Cavtat, Croatia, June 25-28 2007.

[6] He, X., Sun, X-He, Laszewski, G.V., "QoS Guided Min-Min Heuristic for Grid Task Scheduling", Journal of Computer Science and Technology 18(4), pp 442-451, 2003.

[7] Afzal, A., McGough, A.S., Darlington, J., "Capacity planning and scheduling in Grid computing environment", Journal of Future Generation Computer Systems 24, pp. 404-414, 2008.

[8] Yagoubi, B., Slinani, Y., "Task Load Balancing Strategy for Grid Computing", Journal of Computer Science 3 (3), pp. 186-194, 2007.

[9] Elmroth, E., Tordsson, J., "Grid resource broking algorithms enabling advance reservations and resource selection based on performance prediction", Journal of Future Generation Computer Systems 24, pp. 585-593, 2008.

[10] Benjamin Khoo, B.T., Veeravalli, B., Hung, T., Simon See, C.W., "A multi-dimensional scheduling scheme in a Grid computing environment", Journal of Parallel and Distributed Computing 67, pp. 659-673, 2007.

[11] Kovalenko, V.N., Kovalenko, E.I., Koryagin, D.A., Ljubimskii, E.Z., Orlov, A.V., Huhlaev, E.V., "Resource manager for GRID with global job queue and with planning based on local schedules", Journal of Nuclear Instruments and Methods in Physics Research A 502 , pp. 411-414, 2003.

[12] Yamaguchi, T., Takahashi, Y., "A queue management algorithm for fair bandwidths allocation", Journal of Computer Communications 30, pp. 2048-2059, 2007.

[13] Palmer, J., Mitanni, I., "Optimal and heuristic policies for dynamic server allocation", Journal of Parallel and Distributed Computing 65, pp. 1204-1211, 2005.

[14]J. MacLean, "Advance Reservations: State of the Art (draft)," GWD-I, Global Grid Forum (GGF), 2003.

[15] A. G. A. Grama, G. Karypis, V. Kumar, "Introduction to Parallel Computing, second edition," Addison Wesley, 2003.

[16] R. L. a. R. B. S Berson, "An Architecture for Advance Reservations in the Internet," Technical report, USC Information Science Institute, 1998

[17] K. I. Foster, "a distributed resource management architecture that support advance reservation and resource co-allocation," International workshop on QoS, 1999

[18] A. S. U. C. S. K. P. R. Buyaa,"GarQ: An efficient scheduling data structure for advance reservations of grid resources," International Journal of Parallel, Emergent and Distributed Systems, 2009

Energy Efficient Routing Protocol Using Time Series Prediction Based Data Reduction Scheme

Surender Kumar Soni

***Abstract*—**Wireless Sensor Networks (WSNs) are such wireless networks consisting of number of tiny, low cost and low power sensor nodes to monitor some physical phenomenon like temperature, pressure, vibration, seismic events, landslide detection, presence of any object etc. The major limitation in these types of networks is the use of non-rechargeable battery having limited power supply. The main cause of energy consumption in such networks is communication subsystem. So in this paper, time series prediction based data reduction scheme is utilized to develop an energy efficient routing protocol so as to minimize the energy consumption of WSNs. The proposed protocol called Data Transmission Reduction using Prediction (DTRP) to prolong network lifetime via energy efficient data prediction that employs single level grid based clustering. The cluster formation and data transmission in the proposed protocol is same as in LEACH except the usage of data prediction models for reducing the overall number of transmission of information to the base station/sink. Other than this, inter cluster and intra cluster transmissions are reduced which results in reduction in energy consumption of WSNs.

***Index Terms*—** WSN, clustering, sensor networks, prediction model, energy efficient.

I. INTRODUCTION

Based on the WSN architecture and power breakdown, several approaches have to be exploited, even simultaneously, to reduce power consumption in wireless sensor networks. [1][4][6]. Our proposed work in this paper to reduce number of transmissions to maximize WSN lifetime is based on data prediction. Data transmission reduction schemes based on time series grey prediction models (such as GM(1,1) and rolling models) [7][8] have been used to predict the future values of the events in the sensor field so as to develop an energy efficient single hop cluster based protocol.

The proposed protocol results in reduction of numbers of actual transmission of data to the base station/sink and reduction in energy consumption of WSNs. Since transmissions consumes more energy than the processing of information so if the numbers of transmissions are made less by providing the required data through prediction at the base station/sink, lot of saving in energy can be achieved. Therefore, prediction of information, which does not require actual transmission, can reduce the number of transmissions. The proposed protocol works like LEACH (Low Energy Adaptive Clustering Hierarchy) protocol [9] except the application of time series prediction models for acquiring the information at the base station/sink from the CHs [2],[3]. CH compares the predicted value of an event with the actual value of the event received from the normal nodes and if the difference in the values found to be more than the prescribed threshold value, then the actual value of an event is transmitted to the base station/sink to replace the predicted value for the same event at the base station/sink. However, if the difference in both the predicted and actual values of the event falls within the limit, then in that time interval or for that round, the data value is not transmitted to the base station/sink as the same is also predicted at the sink. The same process is applied for all the CHs. This way by predicting future values lots of transmissions are avoided which actually need to be carried out otherwise, thus conserving the WSN energy.

Rest of the paper is summarized as follows. Section II describes the work related system model and proposed algorithm. Section III defines various simulation parameters, performance metrics and explains simulation results. Section IV explains the conclusion of the work.

II. PROPOSED PROTOCOL

In this section, a time series prediction based data transmission reduction have been proposed using GM(1,1) and Rolling models. According to the proposed approach, prediction models have been implemented at the sensor node to predict the future values of data using previous value of each sensor node and at the base station/sink for the same sensor node. Therefore, two queues each for prediction value and for actual value of each sensor node at the respective sensor node and at the base station/sink for each SN have been constructed. The difference between the actual value and the predicted value is called prediction error.

Here, we have defined a maximum value for prediction error called threshold prediction error ϵ. If the difference between the predicted value of information using prediction model and the actual value exceeds thresholds prediction error, then it requires the actual value of SN to be transmitted to the base station/sink. But if the difference between the predicted value using prediction model and the actual value is less than thresholds prediction error ϵ, then it does not require the transmission of information value of that particular SN to the base station/sink. Since the same prediction model is being implemented at the base station/sink, therefore the base station/sink will also assume the same value of energy of that

Surender Kumar Soni is Associate Professor with the Department of Electronics and Communication Engineering, National Institute of Technology, Hamirpur, 177 005 India. E-mail: surender.soni@gmail.com

particular SN without being actually transmitted by the SN. As a result of which the number of actual transmissions are less, therefore the consumption of energy for providing information to the base station/sink is small as compared to the case when no prediction model is applied or when it requires the transmission of information in every round. Definitely higher value of threshold prediction error means less number of actual transmissions and smaller is the energy consumption but compromise has to be made with the accuracy. Larger the value of ϵ, lesser is the number of transmissions and smaller is the energy consumption but lesser is the accuracy as well. Therefore, compromise has to be made between accuracy of data received at the base station/sink and energy consumption. In our proposed work, we have also observed the number of transmissions as a function of threshold prediction error. In cluster based hierarchical routing [1][9]-[11] whole network is divided into clustered layers. Sensor nodes are grouped into clusters with a specific sensor node known as cluster head that performs the task to route the data from the cluster to the base station to form single level clustering. The proposed protocol using time series prediction based data reduction schemes is implemented on single level clustering where it has been assumed that all nodes can transmit with enough power to reach the BS/sink.

We assume that BS/sink has no constraint about its energy resources. Also we assume that BS has total knowledge about the energy level and position of all nodes of the network (most probably by using GPS receiver in each node). The other important assumption of the protocol is random distribution of nodes in network space. The sensor nodes are homogenous, means they have the same processing and communication capabilities and the same amount of energy resources (at the beginning). During data prediction, value of threshold prediction error is fixed at 3%. By simulation, it has been observed that proposed approaches are saving a significant amount of energy in comparison to LEACH protocol (without prediction).

A. System Model

In this research, we assume that set of homogeneous sensor nodes are randomly deployed in the square field to continuously monitor the phenomenon under inspection. The location of the sensors and the base station are set and known apriori. All sensing nodes deployed in the sensing area are assumed to be static and have the knowledge of their location. It is assumed that localization process is carried out just after the deployment of sensor nodes. Nodes are left unattended after deployment; therefore, battery recharge is not possible. All nodes have similar capabilities and equal significance. Each sensor produces some information as it monitors its surrounding area. Generalized energy model based on first order radio energy consumption is used for calculation of energy consumption for sensor nodes within the sensing area [12], [13]. The energy and transmission range of all SNs in the sensor field is such that they all can reach the base station/sink in a single hop. The two models implemented for the prediction of values are grey model (GM(1,1)) and rolling model(RM).

The system model makes the following assumptions:

- There are N sensor nodes that are distributed randomly in d×d square field.
- All the SNs and the base station/sink are stationary after deployment.
- The communication between SNs and base station/sink must be reliable.
- Every SN knows the location of base station.
- Base station/sink is placed at one corner of the sensor field and has enormous energy and memory resources.

B. Algorithm

(i) First, sink sends the broadcast message containing the information about the predefined maximum acceptable prediction error denoted by ϵ.

(ii) Two data queues, prediction value queue, $PVQ_{CH,i}$ at each CH and corresponding $PVQ_{sink,i}$ at the sink node for CH_i is constructed.

(iii) The length of the $PVQ_{CH,i}$ and $PVQ_{sink,i}$ are equal and specified by the same prediction algorithm (i.e. GM (1,1) and RM).

(iv) At the start of every period, $PVQ_{sink,i} = PVQ_{CH,i}$, i.e. both $PVQ_{sink,i}$ and $PVQ_{CH,i}$ store the same values of data of CH_i at sink node and CH respectively.

(v) Actual value of data at any period t is compared with $PVQ_{CH,i}[t]$, the t^{th} item in cluster head prediction queue.

(vi) If the difference between the predicted value of data in $PVQ_{CH,i}$ and actual value of data is greater than ϵ, the $PVQ_{CH,i}[t]$ is replaced with actual value of data and the replaced value of $PVQ_{sensor,i}[t]$ is sent to the sink.

(vii) Sink node will update the t^{th} item in its $PVQ_{sink,i}$ queue. Now the prediction model uses the updated data value for further prediction.

In this way, a number of transmissions of the data between CH and sink nodes can be reduced depending on the accuracy of the prediction model which results in reducing the energy consumption in transmitting the data values to the sink. For predicting future values, time series prediction models such as GM (1, 1) model and RM have been used.

III. PERFORMANCE EVALUATION

In this section we evaluate the performance of proposed algorithm using MATLAB. We first define simulation parameters, performance metrics used and the scenarios created. We then see the effect of various factors like number of nodes and threshold prediction error on evaluation metrics to measure the effectiveness of proposed algorithm. Simulation has been performed for DTRP protocol with GM(1,1) (called DTRP-GM) and DTRP with RM (called DTRP-RM). Simulation results are compared with LEACH [9] approach which is a clustering technique. In this approach no prediction model is applied and the CHs have to transmit the actual information to the base station/sink in every round of its operation.

A. *Simulation Parameters*

We consider a flat and square sized wireless sensor field of size 100 × 100 m^2 in which SNs are randomly deployed. All nodes are homogeneous. Various simulation parameters are listed in Table 1.

B. *Performance Metrics*

Network Lifetime

Network lifetime of wireless sensor network is the time span from the deployment to the instant the network works and is able to achieve its objectives. During our simulations, we have considered following metrics to measure network lifetime:

- FND: number of rounds after which first node dies.
- HND: number of rounds after which 50% nodes die.
- Alive nodes: It is the percentage of alive nodes after specific number of rounds

TABLE 1
SIMULATION PARAMETERS

Parameter	**Default Value**	**Range**
Network size	(100×100) m^2	(50×50) m^2~(400×400) m^2
Number of nodes	400	100~500
Transmission range (r)	100 m	20m ~140 m
Sink location	(0, 0)	
Initial Energy of node	2 Joule	
Data packet size	100 byte	
ε (Prediction Error)	3 %	1~7 %
Window Size	3	
E_{elect}	50 nJ/bit	
ε_{fs}	10 pJ/bit/m^2	
ε_{amp}	0.00134 pJ/bit/m^4	
Data aggregation (E_{DA})	5 nJ/bit/signal	
P (CHs selection probability)	0.05	

Energy Consumption

It is energy consumption per round for the operation of whole network. i.e. energy consumed by a set of nodes is the total sum of energy spent by those nodes in performing required network operation (i.e. transmission, reception, idle state and sensing) during one complete round.

C. *Simulation Experiments*

Clustering, path setup and data dissemination model assumed in Section 2 are used for setting network topology and path setup between SNs to BS through CHs. Different network scenarios are created and energy savings achieved are evaluated for each scenario at overall network level. At cluster level, energy consumed by all active SNs in sensor field is summed up. Similarly at CHs level energy consumed by designated CHs is summed up irrespective of energy consumed by other nodes such as BS/sink. Following basic scenarios are created:

Scenario 1:
Effect of number of sensor nodes

A sensor field of size 100×100 m2, where the transmission range of SNs is kept fixed at 100 m and a numbers of SNs are varied from 100 to 500 in steps. This scenario is created to observe the effect of SN density on network lifetime, network latency and energy consumption at overall network level. This scenario also helps in judging optimum number of SNs within a given sensor field.

Scenario 2:
Effect of number of rounds

For a fixed number of SNs (400), transmission range 100 m and the network size of 100×100 m2, network is operated up to 10000 numbers of rounds and the number of nodes who remained alive has been recorded.

Scenario 3:
Effect of threshold prediction error

For a fixed number of SNs (400), transmission range 100 m and the network size of 100× 100 m2, threshold error is varied to study its effect on the number of transmission of information to the BS/sink.

Effect of number of nodes on network lifetime:

Fig.1 and Fig.2 show the result of network lifetime as a function of number of sensor nodes. Fig.1 gives the variation of network lifetime calculated in terms of number of round spent until the first node in the network dies. Fig 2 indicates the variation of network lifetime obtained in terms of number of rounds spent until 50% node in the network die. To see the effect of the number of nodes on the network lifetime, network size and threshold prediction error is kept fixed at its default values. Simulation results have been compared with LEACH. As can be seen from the graph, network lifetime increases both in terms of FND and HND with increase in the number of nodes.

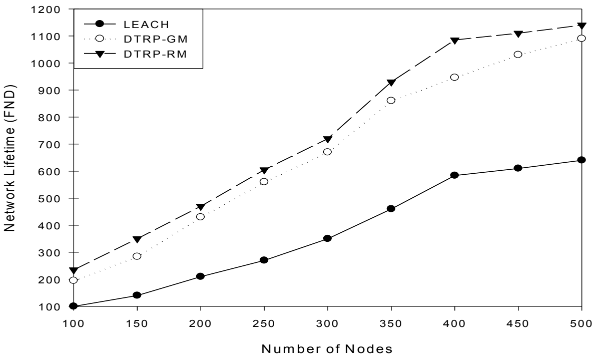


Fig.1. Effect of number of nodes on FND

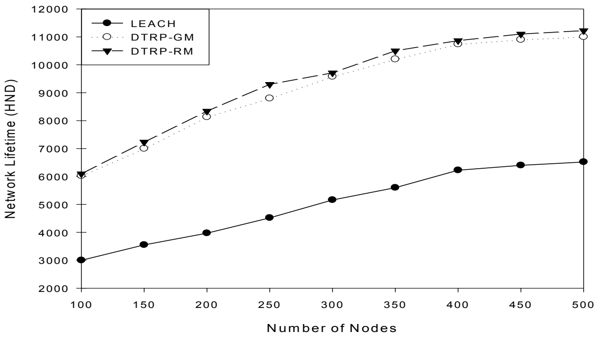


Fig.2. Effect of number of nodes on HND

This is because in all the algorithms, keeping network size fixed number of nodes per cluster increases with increase in node density and the SNs find their turn of transmission after a long time i.e. after more number of rounds.

Hence the number of rounds for the FND and HND to occur are more. If we compare our protocol DTRP with LEACH, results are better in DTRP because in DTRP algorithm, data transmission does not take place if the predicted value differs from an actual value by less than 3%, whereas the data transmission is a continuous process in LEACH. Therefore, the numbers of data transmissions are reduced in the proposed algorithms which ultimately results in saving lot of energy. From the results it has been observed that DTRP-GM algorithm improves the lifetime by a factor of 1.95 while using DTRP-RM an improvement by factor of 2.07 has been observed.

Effect of number of sensor nodes on energy consumption:

Fig 3 shows the result of network energy consumption as a function of node density. Keeping threshold prediction error (ε) and network size constant at its default value, network energy consumption increases with increase in node density. Increasing node density means more number of nodes per cluster. Therefore, increase in number of nodes results in generation of more number of packets by the SNs. This increase in data packets are forwarded to their respective CH for further transmission to the base station/sink. The process up to the cluster head selection, cluster formation, data sensing and forwarding to the CH is same in LEACH, DTRP-GM (1,1) and DTRP-RM. Up to the reception of data by the CH, the energy consumption will remain same in all the algorithms and definitely increases with increase in node density. This is because with more number of nodes more packets will be generated and transmitted. Once the data is received at the CH from the SNs, whole data is further transmitted to the base station/sink directly in single hop but in DTRP-GM and DTRP-RM, prediction algorithm comes in to action. If the threshold limits are satisfied then data is not actually transmitted to the sink and same value is predicted at the sink by the application of same algorithm. Therefore, the number of transmissions is avoided in DTRP which results in increase in energy consumption with increase in node density at a very small rate.

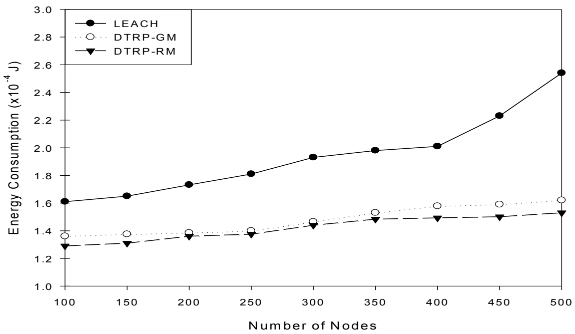


Fig.3. Effect of no of nodes on energy consumption

In other words, energy consumption in DTRP-GM and DTRP-RM increases very slowly because here the number of transmissions using prediction is reduced and data in every round is not required to be transmitted to the BS/sink.

Effect of number of rounds on alive nodes:

Keeping threshold prediction error (ε), node density and network size constant at default values, Fig 4 shows the result of percent nodes alive as a function of number of rounds. From the Fig4 , it is clear that all the nodes are surviving up to approximately 600-700 rounds in LEACH and 1200 rounds in DTRP-GM(1,1) and DTRP-RM.

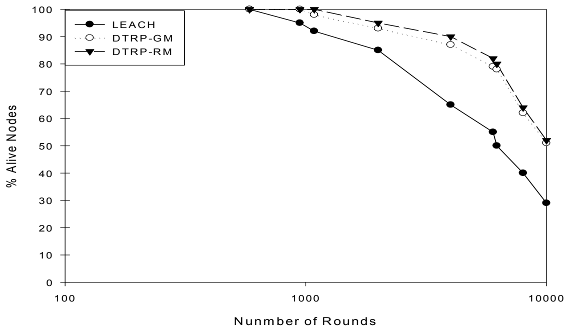


Fig.4. Effect of no of rounds on % alive nodes.

FND in case of LEACH is approximately 700 rounds and in case of DTRP-GM and DTRP-RM about 1200 number of rounds. Thereafter, with the increase in network operations the number of nodes surviving decreases exponentially. After 10000 rounds only 26 percent of nodes remain alive in LEACH and about 52 percent in DTRP-GM and DTRP-RM.

Effect of threshold prediction error ε value on number of transmissions:

Keeping node density, network size and number of rounds constant at default values, Fig 5 shows the percent transmission required at different value of threshold error. In LEACH protocol, the data sensed by the SNs is forwarded to their CH for further transmission to BS/sink. CH in LEACH protocol transmits the data to the sink.

Therefore, the number of transmissions in LEACH have no relation with the threshold prediction error and as shown in Fig 5, the number of transmissions remained constant in LEACH.

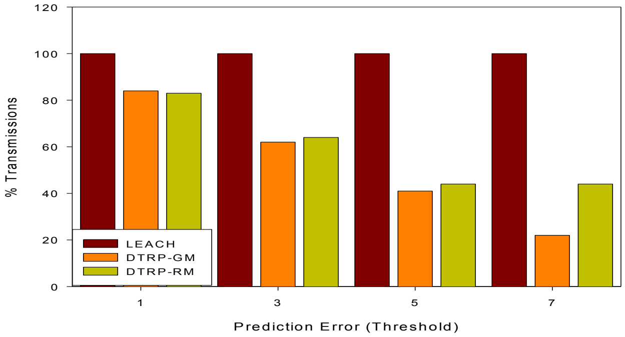


Fig.5. Effect of threshold prediction error on no of tx..

Number of transmission in DTRP-GM and DTRP-RM are inversely related to the threshold limit. Higher the value of ε, lower is the number of transmissions and vice-versa. But increasing value of threshold error leads to inaccuracy in the information received. Saving network energy at the cost of lost accuracy or wrong inference is seriously damaging and unwanted situation for most applications.

IV. CONCLUSION

Two models i.e. GM(1, 1) and RM (rolling model) have been used to reduce the number of sensed data transmissions thus resulting in enhancement of lifetime of WSN. The proposed data prediction models are able to enhance the network lifetime in WSN. In contrast to LEACH where continuous transmission of sensed data consumes more energy, the proposed models reduce number of data transmissions, thus resulting in energy efficiency. An optimal value of window size has been obtained through simulation that leads to lower energy consumption. In contrast to GM(1,1) model, the RM model results in lower computational complexity and requires less buffering to store the past history.

REFERENCES

[1] K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks." Elsevier Journal of Ad Hoc Networks 3 (3), 2005, 325.349.

[2] M. Younis, M. Youssef and K. Arisha, "Energy aware management in cluster-based sensor networks., Computer Networks, 43 (5), 2003, 649. 668.

[3] R. Akl and U. Sawant, "Grid-based Coordinated Routing in Wireless Sensor Networks," IEEE Conference on Consumer, Communications and Networking (CCNC), Las Vegas, pp. 860-864, 2007..

[4] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci, "Wireless Sensor Networks: A Survey," Computer Networks, Vol. 38, No. 4, pp. 393–422, March 2002.

[5] S. Cho and A. Chandrakasan, "Energy-Efficient Protocols for Low Duty Cycle Wireless Micro Sensor," Hawaii International Conference on System Sciences, Maui, HI Vol. 2, pp. 174-185, 2000.

[6] E. Fasolo, M. Rossi, J. Widmer and M. Zorzi, "In-Network Aggregation Techniques for Wireless Sensor Networks: A Survey," IEEE Wireless Communications, Vol. 14, No. 2, pp. 70-87, April 2007.

[7] Ujjwal Kumar and V.K. Jain, "Time Series Models (Grey-Markov, Grey Model with Rolling Mechanism and Singular Spectrum Analysis) to Forecast Energy Consumption in India," Energy, vol. 35, pp. 1709-1716, 2010.

[8] S.K. Madria, Bharat Bhargava, E. Pitoura and Vijay Kumar, "Data Organization Issues for Location-Dependent Queries in Mobile Computing," Springer Verlag Lecture Notes in Computer Science No. 1884, in cooperation with ACM SIGMOD, pp. 142-156, 2000.

[9] W.B. Heinzelman, A.P. Chandrakasan and H. Balakrishnan, "An Application Specific Protocol Architecture for Wireless Microsensor Networks," IEEE Transactions on Wireless Communications, Vol. 1, No. 4, pp. 660–670, 2002.

[10] C. Wang, K. Sohraby, B. Li, M. Daneshmand and Y. Hu, "A Survey of Transport Protocols for Wireless Sensor Networks," IEEE Network, Vol. 20, No. 3, pp. 34–40, 2006.

[11] S.K. Singh, M.P. Singh and D.K. Singh, "Routing Protocols in Wireless Sensor Networks–A Survey" International Journal of Computer Science & Engineering Survey, Vol. 1, No. 2, Nov 2010.

[12] G.J. Pottie and W.J. Kaiser, "Wireless Integrated Network Sensors, " Communications of the ACM, Vol. 43, No. 5, pp. 51–58, May 2000.

[13] J. Chinrungrueng, U. Sununtachaikul and S. Triamlumlerd, "A Vehicular Monitoring System with Power-Efficient Wireless Sensor Networks," ITS Telecommunications Proceedings, pp. 951-954, 2006.

Cloud-based Tele-Monitoring System for Water and Underwater Environments

Georgiana Raluca Tecu, George Suciu
Faculty of Electronics, Telecommunications and IT
University POLITEHNICA of Bucharest
Bucharest, Romania
george@beia.ro

Adelina Ochian, Simona Halunga
Telecommunication Department
University POLITEHNICA of Bucharest
Bucharest, Romania

Abstract— **Recent research in communications and computer science has been considered to advance the performances of monitoring water environments. However, constrains produced by the water environments, caused by the specific channel propagation and harsh operating conditions must be taken into account. The purpose of this paper is to define and describe a monitoring system for the water environments, based on a previous study regarding both the underwater, but also technologies that are appropriate for such surroundings. The system is based on an underwater sensors network which is connected to a cloud platform by means of a reconfigurable wireless transceiver. The sensor network integrates several low cost sensors that can measure different parameters such as water level, the water flow, temperature, pressure etc. The measured parameters will be transmitted through an operational communication node, which should be able to ensure a reliable communication with timing and variation delay constraints. Finally, the paper describes the platform interface available to end users, providing real time visualization of the water environment events.**

Keywords—wireless sensor networks; underwater sensor network; water environment monitoring; cloud computing

I. INTRODUCTION

Recently, there has been an increasing interest in monitoring water environments, the information regarding such surroundings being required by a wide audience, including research scientists, policy-makers, and also the general public. Parameters like water level, flow and sediment data are used by decision makers to resolve issues related to sustainable use, infrastructure planning and water apportionment [1]. Hydrological models use the data to improve the forecasting of floods and water supplies, and to predict the impacts of changes on flow regimes to human and aquatic health and economic activity.

There are several underwater events that can be measured, like the river level or some additional parameters like rainfall, the pressure level or temperature. The purpose of this paper is to present a solution that integrates an underwater network with low cost sensors connected to a cloud platform that can offer real time information. The sensor network integrates several sensors that can be used for different problems and can measure the water level, the water flow, temperature, pressure, but also some parameters that define the water quality. All this information is available in a cloud platform responsible for the collection of environmental data. The platform provides an interface that users can access anywhere, at any moment.

Underwater sensor networks feature a large number of nodes, consisting of static or mobile underwater sensors. These nodes may include pressure sensors used for depth approximation, temperature sensors such as thermistors, photo-diodes for measuring ambient light, fiber optic sensors, cameras and more. Sophisticated but also more expensive sensors may also be used, such as electrochemical sensors for marine environmental monitoring, or acoustic vector sensors that measure the scalar and vector components of the acoustic field in a single point, allowing for a multichannel receiver in a compact form.

A sensor network can offer coverage for a large area, compared to a single instrument platform, which can be an effective sensing tool, but it is limited to take measurements one location at a time. A sensor network can also provide differential measurement, in order to indicate different levels for a certain parameter in different points [2].

Furthermore, cloud computing provides a platform for processing big data from hundreds of different sensors, enabling the analysis of the environmental data through a large sample of datasets [3].

The paper is organized as follows: Section II describes the rationale for monitoring water environments considering the state of the art and strategic relevance, while Section III presents challenges of WSNs. Section IV presents the proposed tele-monitoring system and describes the components. Finally, Section V concludes the paper.

II. RATIONALE FOR WATER AND UNDERWATER MONITORING SYSTEMS

Despite the fact that significant progress has been made regarding sensors, communication and computing technology, the underwater sensor platforms are generally inferior to their terrestrial counterparts [4]. In opposition to terrestrial Wireless Sensor Networks (WSNs), underwater sensor nodes are more expensive and there are developed fewer sensor nodes. Data acquisition from the sensor nodes is made through autonomous underwater vehicles [5]. Also, compared to a dense deployment of sensor nodes in a terrestrial network, a sparse deployment of sensors is placed underwater [6].

The communication in the underwater environments is wireless and is established through the transmission of acoustic waves. The general problems are the limited bandwidth, long propagation delay and also the signal fading issues. Another challenge is the sensor node failure, due to the harsh environmental conditions [7]. These sensors must be able to perform self-configuration and calibration, and also they have to adapt to these environment conditions. The issue of energy conservation for underwater networks involves developing efficient underwater communication and networking techniques [8].

An Underwater Sensor Network offers a different vision, providing a real time visualization of the underwater events. The existing wireless sensor networks and tools for the development of such infrastructures are currently hot topics under research for their applicability in underwater scenarios. There are several projects and devices that offer solutions, from water level monitoring sensors used in agriculture to systems deployed on rivers banks, which aggregate multi-parameter sensors or various combinations of individual sensors.

The proposed solution consists in an underwater network with low cost sensors connected to a cloud platform that can offer real time information. The sensor network integrates several sensors that can be used for different problems and can measure the water level, the water flow, temperature, pressure, but also some parameters that define the water quality. All this information is available in a cloud platform responsible for the collection of environmental data. The platform provides an interface that users can access anywhere, at any moment, via Internet.

Existing sensors in the Black Sea's offshore area detect environmental change events and provide early warning messages—essential information for emergency organizations and residents. Furthermore, Danube's earthquake monitoring system provides data that is helping to minimize risks to residents and property. Monitoring the Danube underwater delta reveals slope stability changes that are of significance to the nearby coal port, container terminal and Zimnicea ferry terminal [9].

By using weather stations with online tele-monitoring capabilities of water parameters, new coastal monitoring services will report live weather and sea states status in the Black Sea, helping all vessels make safe decisions. Continuous observations at key sites in coastal Black Sea are helping to address global challenges, manage marine resources, monitor regional environmental and climate change, and detect hazards to coastal communities.

Underwater monitoring quantifies changes in marine environments and impacts on resources, helping shape sustainable resource management. The monitoring of major marine currents and ecosystems help decision-makers understand and predict the consequences of a spill and allow for effective critical response planning [10].

River environment monitoring helps track river productivity, including the growth and impact of phytoplankton blooms, key to productive fisheries and to mitigating the commercial and health effects of harmful red tides. Acoustic monitoring in both networks shows strong promise for further improvements in measuring fish stocks and predicting returns.

III. Technical Challenges of WSNs for Tele-Monitoring Water Environments

Pollution monitoring, harbour surveillance, undersea archaeology, river bottom seismic research, river life observation are among some of the fields that can benefit from the wide opportunities that Underwater Sensor Networks (UWSNs) offer. Nevertheless, before UWSNs become commercially available or widely used, there are certain issues to be addressed, as presented is this section.

A. Technological challenges

Localization is one of the major and challenging tasks in UWSNs [11]. It is important because raw sensor data without spatio- temporal tagging does not provide much information. It is challenging because GPS signal does not propagate through water and alternative cooperative positioning schemes are not applicable in practice due to acoustic channel properties. Acoustic channel have low bandwidth, high propagation delay and high bit error rate. The speed of sound is approximately 1500 m/s, yet it varies with temperature, pressure and salinity.

Another challenge is the mobility. Moreover, energy limitation is still an issue as it is in other sensor networks.

Because we intend to develop an optimal solution that integrates several fields, and to deploy this solution worldwide, we have to take in account all the requirements and limitations and adapt our solution to different environments and conditions. To solve this problematic, we need to conduct a state of art of all the equipments that are already deployed.

The main innovation represents the integration of a wireless underwater network, representing a solution that offers a low cost and low power wireless sensor network that efficiently uses available energy without compromising performances (range, data rate, latency, standard compliance). The main challenge is to use different types of low power sensors for implementing this wireless network.

B. Applicability of Standards

WSNs comprise of a large number of spatially distributed autonomous devices that may collect data using a wireless medium. They may be used to cooperatively control and monitor physical or environmental conditions, such as temperature, pH, electrical conductivity salinity, chlorophyll, sound, vibration, pressure, motion or pollutants, at different locations [12]. International standards for wireless devices and networks, such as ZigBee, WirelessHART and ISAIOO.Il a use stacks to provide a layered and abstract description of the network protocol design [13], [14]. Each layer in the stack is a collection of related functions, and each layer is responsible for providing services to the layer above it, while receiving services from the layer below it [15], [16], [17].

IV. Proposed Tele-Monitoring System

This section presents the main components of the tele-monitoring system for water environments. In Fig. 1. we present the proposed network architecture for a UWSN composed by a low power sensor, a transceiver and an access point that provides the collected data to a cloud platform.

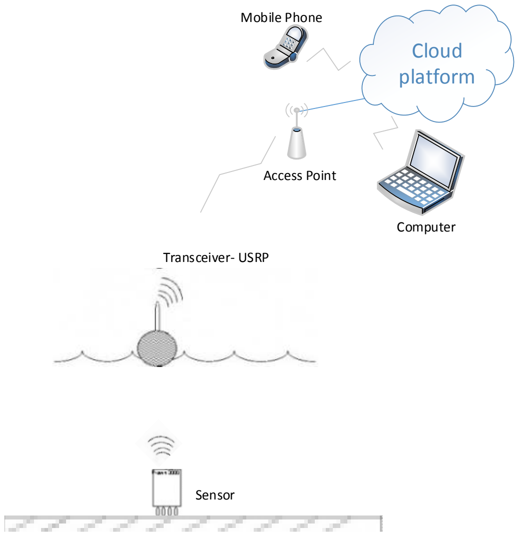


Fig. 1. Network architecture for UWSN

A. Underwater Sensors

The underwater sensor is wireless connected to the transceiver, which is powered by a local battery and solar panel, thus providing a long period of working autonomy for the system. Moreover, the transceiver stores the data and transmits at programmed intervals or when thresholds are reached in order to save battery power via GPRS or UHF unlicensed band, when the GSM signal is not available. The main environmental parameters measured by the underwater sensors are presented in Fig. 2.

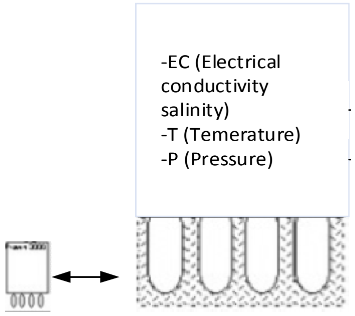


Fig. 2. Underwater Sensor

B. Transceiver and Access Point USRP

The USRP is a hardware platform that can be interfaced to a miniaturized host computer via a USB port in order to be used to create a real-time software defined radio (SDR) utilizing open-source GNU radio software [18]. It is a motherboard which contains a Field Programmable Gate Array (FPGA) for high-speed signal processing, 4 high-speed analog-to-digital converters (ADC), 4 high-speed digital-to-analog converters (DAC), and auxiliary analog and digital input/output (IO) ports.

C. Software and Cloud Platform

We used GNU Radio because it is a free software development toolkit that provides signal processing to implement SDRs using off-the-shelf RF hardware [19]. GNU Radio applications are primarily written using the Python programming language, while the performance-critical signal processing tasks are implemented in C++, as presented in Fig.3.

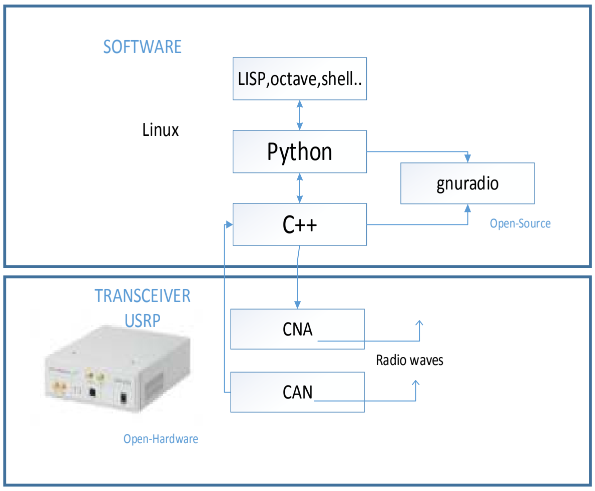


Fig. 3. Transceiver Hardware and Software integration

The platform was based on the Ubuntu Linux – Apache – MySQL software releases using our SlapOS decentralized Cloud platform hosted on several server nodes. The architecture is based on the concept of Master and Slave nodes.

Master nodes are central directory nodes cloud system, serving to allocate processes to Slave nodes and keep track of the situation of each slave node and software that are installed on each node. Slave nodes can be installed on any computer, both in data centers and in private networks and their role is to install and run software processes.

Slave nodes request to Master nodes which software they should install, which software they show run and report to Master node how much resources each running software has been using for a certain period of time. Master nodes keep track of available slave node capacity and available software. Master node also acts as a Web portal and Web service so that end users and software bots can request software instances which are instantiated and run on Slave nodes as computer partitions.

V. Results and Discussions

The proposed system provides several parameters like water level, flow and sediment data. The measured parameters are provided in a user interface which also allows a graphical representation of the events.

To verify the measurements of the underwater sensor we compared the values captured by the underwater pressure sensor to determine the level of the water in a lake with the values of a level sensor located at the surface of the water.

First, the precipitation intensity in mm rain/sqm was measured during two months and was considered negligible, due to a drought period, as seen in Fig. 4.

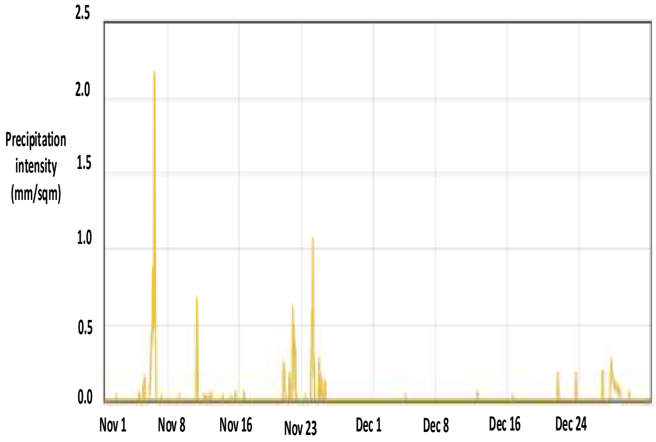


Fig. 4. Precipitation intensity graph

Next, the water intake of the river entering the lake was calculated and the lake level was measured by the underwater sensor, as seen in Fig. 5.

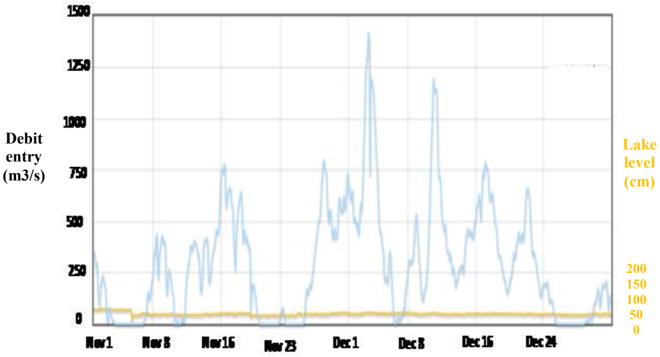


Fig. 5. Lake level in cm measured by underwater sensor (orange), surface level sensors (grey) and debit entry in m3/s (blue)intensity graph

It can be observed that the lake level is not affected by the water intake of the river, as both underwater and surface level sensor indicated.

VI. Conclusions

In this paper we presented the main challenges for a tele-monitoring UWSN system and elaborated the specific requirements for an innovative implementation and analyzed the strategic relevance of such a system. We proposed a system for monitoring the water level by using two different methods, presenting the main components and measurement results.

As future work we envision to develop a cloud platform for aggregating the data gathered from the access points, thus providing higher availability and distributed processing capacity. We could accomplish the implementation and the deployment as a Cloud Service for MAPE (Monitoring, Analyzing, Planning and Enforcing).

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/134398.

References

[1] S. B. T. Sany, R. Hashim, M. Rezayi, A. Salleh, and O. Safari, "A review of strategies to monitor water and sediment quality for a sustainability assessment of marine environment," in Environmental Science and Pollution Research 21, no. 2, 2014, pp. 813-833.

[2] G. Suciu, S. Halunga, O. Fratu, A. Vasilescu, and V. Suciu, "Study for renewable energy telemetry using a decentralized cloud M2M system," in Wireless Personal Multimedia Communications (WPMC), IEEE 16th International Symposium on, 2013, pp. 1-5.

[3] A. Sharma, and U. C. Vinayak. "Efficient Data Storage in Cloud." In IJITR 2, no. 3, 2014, pp. 977-980.

[4] M. Felemban, and E. Felemban, "Energy-delay tradeoffs for Underwater Acoustic Sensor Networks," in Communications and Networking (BlackSeaCom), IEEE First International Black Sea Conference on, 2013, pp. 45-49.

[5] A. Speers, A. Topol, J. Zacher, R. Codd-Downey, B. Verzijlenberg, and M. Jenkin. "Monitoring underwater sensors with an amphibious robot." In Computer and Robot Vision (CRV), IEEE Canadian Conference on, 2011, pp. 153-159.

[6] R. Kastner, A. Lin, C. Schurgers, J. Jaffe, P. Franks, and B. S. Stewart, "Sensor platforms for multimodal underwater monitoring," in IEEE International Green Computing Conference (IGCC), 2012, pp. 1-7.

[7] N. Srivastava, “Challenges of Next-Generation Wireless Sensor Networks and its impact on Society”, in Journal of Telecommunications, Volume 1, Issue 1, 2010, pp. 128-133.

[8] J. H. Cui, J. Kong, M. Gerla, and S. Zhou. "The challenges of building mobile underwater wireless networks for aquatic applications," in Network, IEEE 20, no. 3, 2006, pp. 12-18.

[9] I. Ralita, A. Manea, “The monitoring of risk meteorological phenomena in real time, by means of NIMS applications,” in Geographical Phorum – Geographical studies and environment protection research, Year 6, No. 6, 2007, pp. 121-126

[10] A. Davis, and H. Chang, "Underwater wireless sensor networks." In Oceans, IEEE, 2012, pp. 1-5.

[11] M. Erol, L. F. M. Vieira, and M. Gerla, "Localization with Dive'N'Rise (DNR) beacons for underwater acoustic sensor networks," in Proceedings of the second workshop on Underwater networks, ACM, 2007, pp. 97-100.

[12] R. Su, R. Venkatesan, and C. Li. "A new node coordination scheme for data gathering in underwater acoustic sensor networks using autonomous underwater vehicle." In Wireless Communications and Networking Conference (WCNC), IEEE, 2013, pp. 4370-4374.

[13] W. Ping, F. Donghao, X. Jianchun, Y. Qiliang, W. Ronghao, and W. Wenhao, "An improved MAC protocol for underwater acoustic networks," in Control and Decision Conference (CCDC), 2013 25th Chinese, IEEE, 2013, pp. 2897-2903.

[14] G. Fan, H. Chen, L. Xie, and J.H. Cui, "A bidirectional TDMA protocol for underwater acoustic sensor networks," In Proceedings of the Eighth ACM International Conference on Underwater Networks and Systems, ACM, 2013, p. 12.

[15] W.H. Liao, and C.C. Huang, "SF-MAC: A spatially fair MAC protocol for underwater acoustic sensor networks," in Sensors Journal, IEEE 12, no. 6, 2012, pp. 1686-1694.

[16] P. Pandey, M. Hajimirsadeghi, and D. Pompili, "Region of Feasibility of Interference Alignment in Underwater Sensor Networks," 2011, pp. 1-14,

[17] A. Vulpe, S. Obreja, and O. Fratu, "Interoperability procedures between access technologies using IEEE 802.21," in Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE), IEEE 2nd International Conference on, 2011, pp. 1-5.

[18] GNU Radio [Online – Accessed June 2014]: www.gnuradio.org

[19] A. Martian, L. Petrica, and O. Radu. "Cognitive radio testing framework based on USRP." in Telecommunications Forum (TELFOR), IEEE, pp. 212-215.

Detection and prevention from denial of service attacks (DoS) and distributed denial of service attacks (DDoS)

Nozar kiani , Dr. Ebrahim Behrozian Nejad

Institute For Higher Education ACECR Kouzestan , Iran

Kiani.nozar@gmail.com

Abstract—regarding the growing trend of denial of service attacks (DoS) and distributed denial of service attacks (DDoS) in the context of internet networks, and the importance of Web-based services in these networks, we need to be quite aware of these attacks. Although it is difficult to study these attacks, through having a good insight about the effects and consequences of these attacks, it is possible to obtain the preventative ways for these kinds of attacks in order not to provide a necessary context for aggressors of these kinds of attacks. And the servers provide their services properly, and the users get the resources and services without any disruption. Although the prediction and deviation of these attacks in a wide area like web in a global scale is difficult, we can handle these attacks using some preventative techniques in the context of network, and detection of attack operations and the deviation of attack during the attack to reduce the effects of attack. Unfortunately, with the enormous traffics of attacks some damages have been found . Thus, detection of attacks DDOS At the earliest possible time is more favorable than waiting for the spread of a comprehensive flood of attacks. To implement an efficient defense system, we should use a network topology leverage to monitor the distributed traffic and detection. In this study, the Preventative methods for these attacks will be explained.

Keywords— DOS Attack, D.DOS Attack, Stacheldraht Attack, SYN flood, Legitimacy testing , Traceback, Trinoo Attack

1 - Introduction

The purpose of the DOS attacks is to Interfere with resources and services that users are going to access and use them (disabling the services.) The main purpose of these kinds of attacks is to prevent users from accessing to a particular resource. In These attacks, attackers using several techniques make attempt to put into trouble the authorized users to access and use a particular service, and disturb the services of a network. Trying to generate False traffic in the network, interfering with communication between two machines, preventing authorized users from accessing a service, and disrupting services are some instances of other objectives that attackers pursue. In some cases, in order to carry out massive attacks using DOS attacks as a starting point an ancillary element is used to provide a context for the original invasion. Accurate and legitimate use of some resources may also leads to a kind of DOS attacks. A flood of large enough traffic causes to overflow a buffer connections, disk fatigue or saturation of connecting link and so on lead to the crash of the suffered device. And given that in recent years the widespread attacks DDOS is increasing for the competitiveness of business enterprises, service provider sites, and so forth has been conducted. Massive service attacks denial is considered as the greatest threat, therefore, to prevent these growing attacks, some preventative methods will be presented.

1-1 - The Internet constitutes are consumable and limited.

Infrastructure systems and connected networks that make internet are composed of entirely limited resources. Bandwidth, processing power and storage capacity all are limited and the target of common DOS attacks. Attackers perform the attacks trying to consume a significant amount of available resources so that some extent of the services will be disrupted. Abundant resources that have been designed and used properly, may contribute to reducing the impact of an attack DOS , but today's attack methods and tools operate even in the most abundant sources and make interferes in them.

1-2 - Internet security is largely dependent on all the factors.

The DOS Attacks Usually occur from one or more points invisible to the victim's system or network. In many cases, the starting point of the attack includes one or more systems that are provided to an attacker through security exploits, and so the attacks are not done by the system or the piercing systems. Therefore, defense against penetration not only protect the

Internet-related property, but also helps avoid using this property to attack other networks and systems. Then no matter how much your system is protected, exposure to many types of attacks, particularly DOS , Depends largely on the security situation in other parts of the Internet.

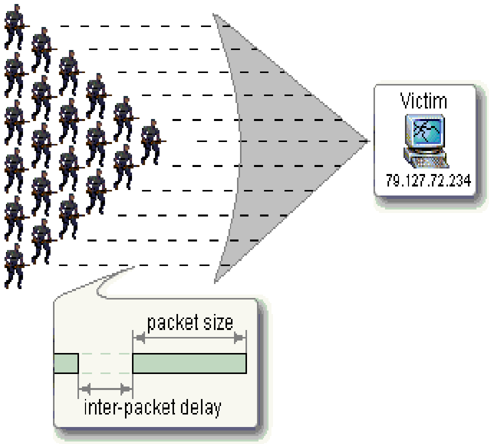


Fig.1 the attack of packets diagram

Defending DOS attacks is not only a practical discussion. Limiting demand amount, packet filtering and manipulation of software parameters can sometimes help limit the effects of DOS attacks provided that the DOS attack is not using all the existing resources. In most cases, we can only have one defensive reaction, and this happens only and only if the source or sources of the attack are determined. Using IP addresses faking during the attack, the advent of distributed attack methods, and existing tools cause a constant challenge against those who respond to DOS attack.

Initial DOS Attacks technology consisted of a simple tool to generate the packages and send them from "one source to one destination". With the passage of time, the tools have progressed toward the implementation of attacks from "a source to several destinations", from "several sources to single destination ", and from "multiple sources to multiple destinations ".

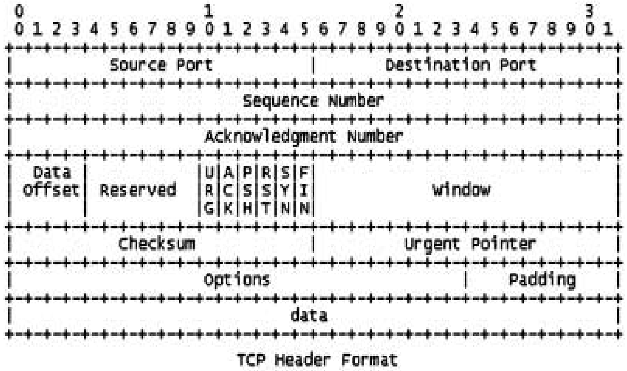
Today, most of the reported attacks to CERT / CC are based on the sending of a very large number of packets to a destination which consequently creates a lot of endpoints and consumes the network bandwidth. Such attacks are typically referred as **Packet flooding**. But about the "attack to multiple targets" fewer reports have been received. [3]

2 - Examining the TCP packets and how to communicate under the TCP / IP protocol

For closer examination and explanation of how DOS attacks function we need to investigate TCP packets and explain how to communicate under the TCP / IP protocol. They will be discussed as follows:

2-1 - Examining the components inside a TCP packet

The internal components of a TCP packet are: Source port, destination port, the data string and so on. They make the information on the way to the internet be displaced .



```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Source Port          |       Destination Port        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Sequence Number                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Acknowledgment Number                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Data |           |U|A|P|R|S|F|                               |
| Offset| Reserved  |R|C|S|S|Y|I|            Window             |
|       |           |G|K|H|T|N|N|                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Checksum            |         Urgent Pointer        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Options                    |    Padding    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             data                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

                       TCP Header Format
```

Fig.2 TCP Packet Format

2-2 -The examination of TCP Protocol Function

In the following Fig, the server named TCP B and the client named TCP A are shown:

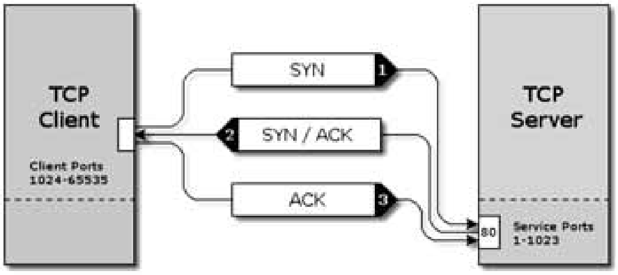


Fig.3 Diagram of connections in TCP

1. The client sends TCP Packet to the server marked with SYN. This packet makes the server realize the client is going to send the information. Then the client is waits for a response to receive and accordingly sends the information .
2. After receiving the client request, the server, in response to the client, sends a packet marked with SYN / ACK indicating the permission to communicate and transmit data.
3. The client sends an ACK to the server after receiving a packet from the server.
4. Then the client tries to send data. [1]

3 - Examination of various DOS methods

3-1 - SYN flood attack investigation

This attack sends numerous requests marked with SYN to the victim machine making Backlog queue full. But, what is Backlog? All requests that enter the machine including SYN mark for Communications. They are stored in order in a part of the memory to be considered and accordingly being answered so that the communication can happen. This part of the memory is called Backlog Queue. When this part is filled with many requests, the server is forced to abandon new requests and as a result, these new requests can't be processed and investigated.

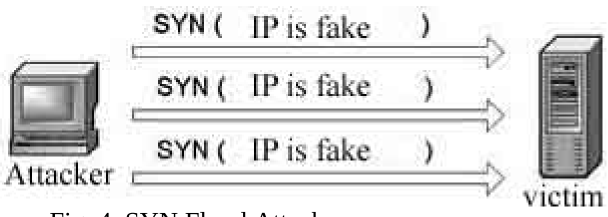


Fig .4 SYN Flood Attack

3-2 - Reset (RST)

Packets which are sent with RST mark cause the connection to be disconnected. In fact, if the machine A sends a packet marked RST to the machine B, the connection request from the Backlog will be cleared.

This attack can be used to disconnect the two machines. That is , The attacker breaks off the established connection between the two machines A And B by sending an RST request to the Machine B from the machine A. in fact, inside the packet sent to the victim from the attacker's machine, IP client is put, and consequently the machine B , which is The server , eliminates the machine A From the Backlog.

In this method, the attacker through using a tool can fake the IP and in fact, sends his request instead of another machine. This technique is also called Spoofing. Fig ()

paying a little attention to Fig 5-1, you will find that Source IP which in the transferred packet sent by the attacker machine to machine B is the same with IP Machine Number A (1.1.1.1 9) , while the IP Machine Number C that the attacker uses is quite another. (1.1.1.3 ([1]

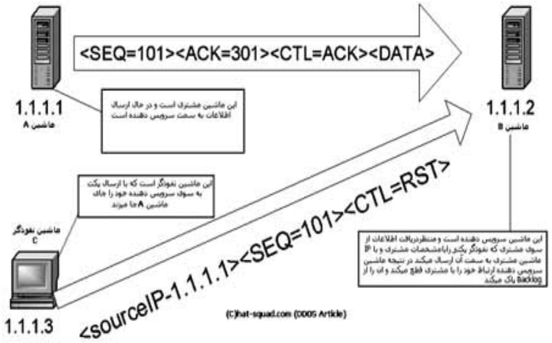

<SEQ=101><ACK=301><CTL=ACK><DATA>
1.1.1.1
ماشین A
این ماشین مشترک است و در حال ارسال اطلاعات به سمت سرویس دهنده است
1.1.1.2
ماشین B
این ماشین نفوذگر است که با ارسال یک به سوی سرویس دهنده خود را جای ماشین A جا میزند
ماشین نفوذگر C
این ماشین سرویس دهنده است و منتظر دریافت اطلاعات از سوی مشترک که نفوذگر یکی را با مشخصات مشترک و با IP ماشین مشترک به سمت آن ارسال میکند در نتیجه ماشین سرویس دهنده ارتباط خود را با مشترک قطع میکند و آن را از Backlog پاک میکند
1.1.1.3
<sourceIP-1.1.1.1><SEQ=101><CTL=RST>
(C)hat-squad.com (DDOS Article)

Fig .5 Attacking RST Attack

3-3 - **Land Attack**

In this attack, using Spoofing method in the packets sent to the server, instead of IP and the port of Source and destination, IP and the port of server's machine is placed. In fact, IP and the port of server's machine are sent to the server. As a result, in the old operating systems an internal loop or Routing appears which consequently fill the memory and gives rise to DOS attack.

In addition, This attack in Win 95 (winsok 1.0) and Cisco IOS ver 10.x machines and the old system makes the system break down, but today all intelligent systems such as IDS are able to identify these attacks and therefore, these attacks do not have any major effect on these server's function.

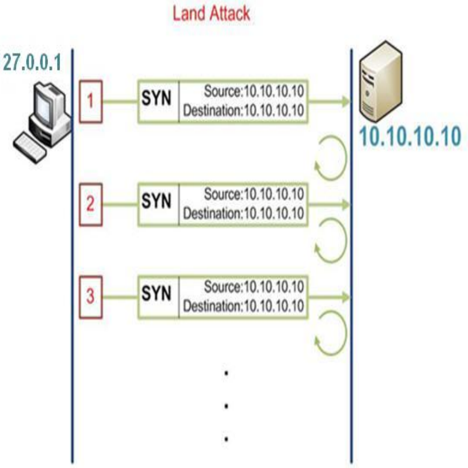


Fig .6 Land Attack [2]

3-4 - Smurf Attack

These attacks by sending ICMP requests to a range of amplifier IPs give rise to traffic extension; this in turn leads to DOS attack.

Attacker can send their ICMP request in a Spoof- like manner and through the victim's machine to the IPs Of amplifier. By sending a request, hundreds of responses to the ICMP request will flow to the victim machine and this raises the traffic (Fig, 6-1).

- **Amplifier**: All networks that have not filtered the ICMP requests for IP broadcast are considered as Amplifier.

the attacker can send some requests to, for example, IPs Such as: 192.168.0.xxx The X can be 255, 223, 191, 15, 9, 127, 95, 63, 31, 15, 7 3 , namely the IPs Of Broadcast . However, it is noteworthy that IP broadcast depends on how IP segmentation in the network is. [1]

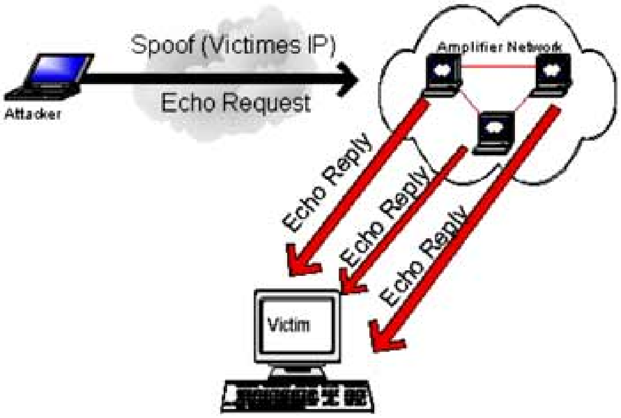


Fig .7 Smurf Attack

3-5 - **Ping Flood or Ping of death**

In this type of attack by a direct request (Ping) to the victim computer, the attacker tries to block the service or reduce its activity. In this type of attack the size of information packets becomes to a great extent (above K64, that is unauthorized in Ping) large and the victim's computer is not able to deal effectively with the mixing packets and it will break down.

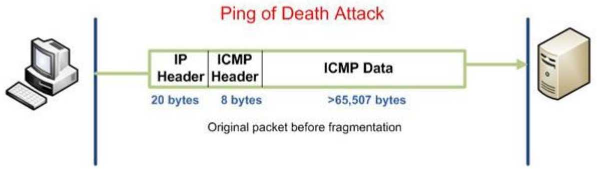


Fig .8 Standard Format Ping [2]

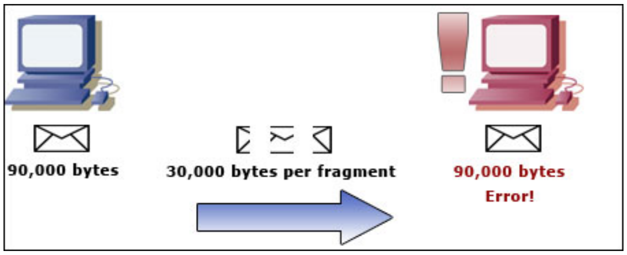


Fig. 9 Diagram of Ping of death attack

3-6 - Teardrop Attacks

When information is transferred from one system to another system, it will be divided into small pieces, and in the destination system, these pieces attach together and become the whole. Each packet contains an offset field, which shows that the packet contains what piece of information. This field, along with the order number helps the destination system to connect the packets again. If the packets are sent with the irrelevant offset number and order, it makes destination system unable to sort them and the system will break.

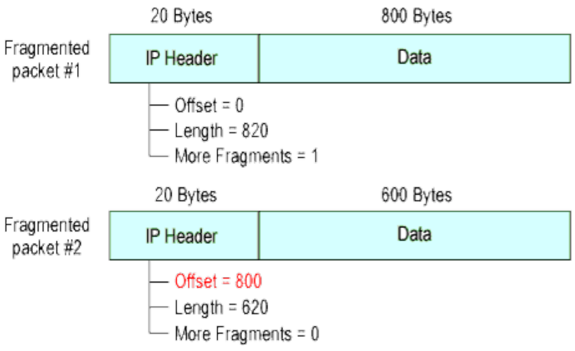


Fig .10 Teardrop Attacks [2]

4 - Distributed Denial of Service D.DOS attacks and various Types of D.DOS attacks

DDOS (Distributed Denial of Service) attacks are kinds of wide distributed DOS attacks. Generally, DDOS is s an organized attack against the available services on the Internet. In this way, DOS attacks are indirectly done on the victim's computer by a large number of hacked computers. The targeted services and resources are called the "Primary victims" and the computers used for the attack are the " Secondary victims ". DDOS attacks are generally more effective in knocking down (disabling) the large companies as compared with DOS attacks.

This type of attack connects the nature of distributive internet with the hosts which have the   separate essence around the world in order to create giant unidirectional flow of packets against one or several victims. To run a DDOS unidirectional flow, hacker first gains the control of a large number of victim devices which is called Zombies.

Zombie systems are placed everywhere in the internet and have a simple vulnerable series that allows hacker to gain the control of system quickly. Till now in these kinds of attacks, Zombie has been installed in vulnerable university servers, the system of large companies, and the system of servers and even in household systems which connect to Loop Digital-Subscriber or Cable Modem services. Hacker scans the large strips of internet to find the vulnerable systems, use them and install Zombies on them. Most of the devices, on which the Zombies is installed, through using the attack of overfilling Buffer mass or a damaging software are installed . Hackers generate hundreds and thousands of Zombies.

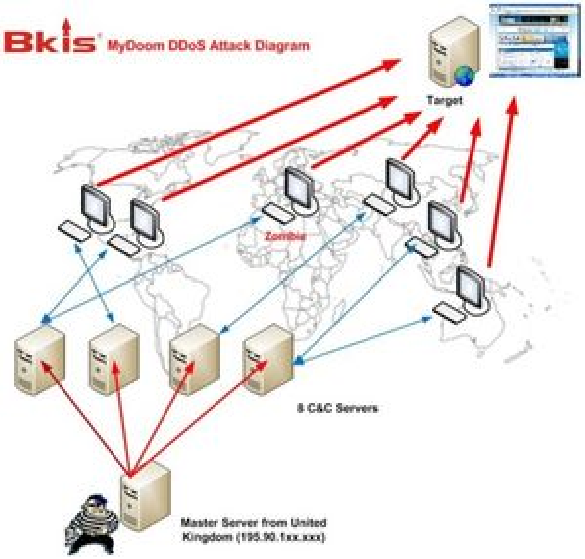

Bkis MyDoom DDoS Attack Diagram
Target
Zombie
8 C&C Servers
Master Server from United Kingdom (195.90.1xx.xxx)

Fig .11 Diagram of attacks D.DOS [3]

Based on the intensity of attacks DDOS Attacks are divided into two categories: disruptive attacks and degrading attacks. In disruptive attacks, providing services from the victim machine to the customers are completely impeded [6]. These attacks in their own turn are divided into three categories: Self-Recoverable, Human-Recoverable and Non-Recoverable. In the first one, namely Self-Recoverable, the victim machine a short while after the attack cease can be recovered automatically. UDP flood And TCP flood attacks fall into this category. In the second type, the system can not automatically recover and requires human intervention. Attacks that lead to rebooting, disabling or capping off the system fall into this category. The third type attacks cause permanent damages to the target system and the retrieval of the system requires purchasing new hardware [9]. [6]

In degrading attacks the purpose of attack is to use some of victim's machine resources. As a result, this causes the delay in attack detection and consequently gives rise to huge damages to the victim machine [5].

Below some instances of Distributed Denial of service attacks D.DOS are Introduced and how the attacks function are explained.

4-1 - Trinoo Attacks

Trinoo is originally a kind of Master / Slave programs that cooperate and synchronize with each other in order to have a flood attack UDP Against the victim's computer are. In a normal process, the following steps occur to establish a Trinoo DDOS network.

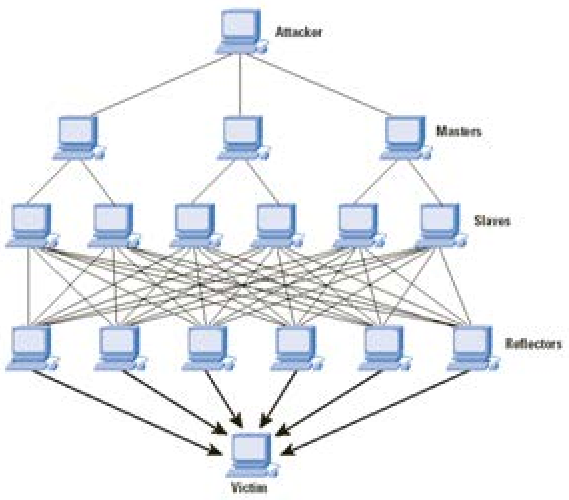


Fig .12 Diagram of Trinoo attack

Step 1: The attacker, using a hacked host, collects a list of systems that can be hacked. Most of this process is done automatically by the hacked host. This host keeps in itself some information including how to find other hosts for hacking.

Step 2: Once this list is ready, the scripts for hacking and changing them into Masters or demons are implemented. A Master can control several Demons. Demons are the hacked hosts that perform the main UDP flood on the victim's machine.

Step 3: DDOS attack is done when command is sent to the hosts of the Master from the attacker. . These masters can command any Demon to have a DOS attack against IP address specified in the command to start and trough doing a lot of DOS attack a DDOS attack Forms [6] [4].

4-2 - TFN/TFN2K attacks

TFN (Tribal Flood Network) is generally a Master / Slave attack in which coordination takes place to have a SYN flooding against the victim's system. TFN demons are able to do much more varied attacks include ICMP flooding, SYN flooding, and Smurf attacks. Therefore, TFN is more complicated as compared with Trinoo attack.

Compared with the main TFN tool, TFN2K has several key advantages and improvements. TFN2K attacks are implemented by faking IP addresses that makes it more difficult to discover the source of the attack. TFN2K attacks are not just simple TFN flood. They also include the attacks that exploit the security gaps of the operating system for invalid and incomplete packets in order to cause the failure of victim systems. TFN2K attackers do not need to run the commands by entering to the Client machine instead of Master in TFN, and they can run these commands from a far distance. The connection between Clients And Demons is no longer restricted to ICMP Echo responses can be done through different intermediaries like TCP And UDP . Therefore, TFN2K are more dangerous and are more difficult to discover as well.

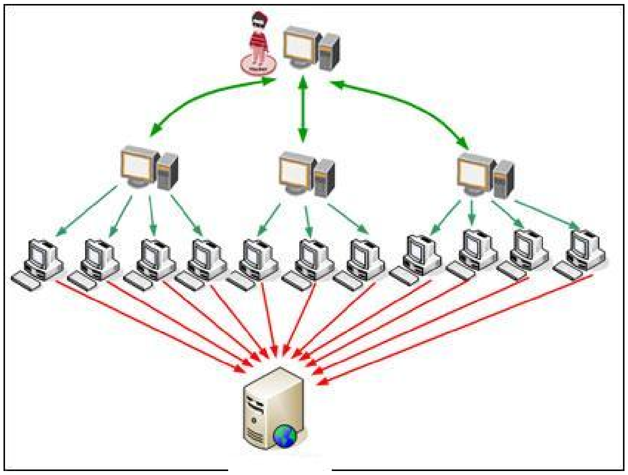


Fig .13 Diagram of TFN/TFN2K attacks

4-3 - Stacheldraht attacks

Stacheldraht code is very similar to Terrinoo and TFN, however, Stacheldraht Allows the communication between the attacker and Master (Which in this attack is called Handler) to be encrypted; the operations can upgrade their code automatically, and they can proceed to do various types of attacks, such as ICMP floods, UDP floods , and SYN floods .

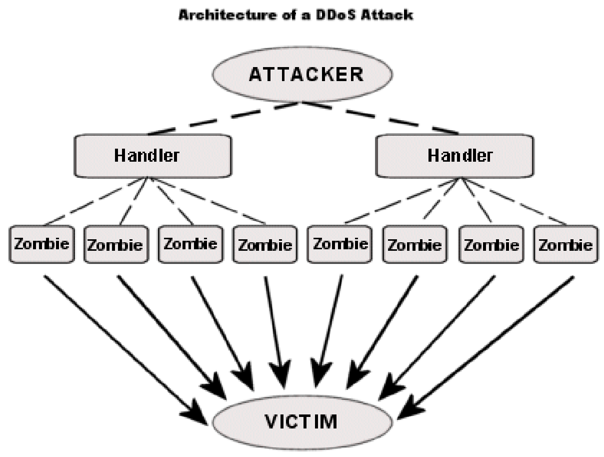


Fig .14 Diagram of Stacheldraht attacks [4]

5 - An example of a DDOS attack

In recent years, DDOS attacks on the Internet have targeted the accessibility. The first case happened on 7 February 2000. In that attack, Yahoo was targeted in a way that its portal was inaccessible for three hours. On February 8, 2000, some Sites like Amazon, Buy.com, CNN and eBay were targeted by the attackers. This gives rise to the complete cancellation of their operations or makes them slow down considerably. According to published reports, within the 3 hours that Yahoo was attacked the Commercial and advertising benefit amount that was lost was about 500, 000 dollars. According to the statistics provided by Amazon, within the 10-hour that this site was attacked 600, 000 dollars have been lost. Furthermore, During the DDOS attack accessibility amount of Buy.com was reduced from 100% to 9.4% and the users' volume of CNN has been lowered and became 5%.

DDOS attacks are more powerful and more difficult to detect and cop with as compare with DOS attacks . The reason is that in these attacks several machines can coordinate in order to send a small stream of traffic to the target machine and the control of all the traffics is hard for the target machine [4].

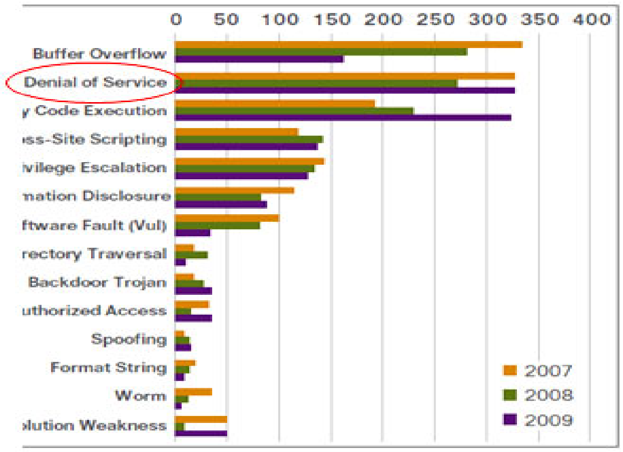


Fig .15 the important threats, vulnerabilities of computer systems.

6 - Ways of Coping

6-1 - Defense against Smurf attacks

If you are exposed to the Smurf attack, you can't do anything special. Although this is possible to block the attacker packets in the external router, the origin of the source width band of the router will be blocked. In order for the network provider above you to the attacks at the source of attack, the coordination is needed.

In order to prevent the attack from your site, your external router should be conFigd in a way that blocks all the outgoing packets that have a source address inconsistent with your subnet. If the faking packet (the packet which does the action of faking) can't go out, it cannot make a serious damage.

To avoid being as an intermediary and participating in other person's DOS attack , conFig your router in a way that block the packets which their destination is all addresses of your network. That is to say, do not allow the ICMP released packet on your network to come to the router. It allows you to have the ability to keep performing the action of ping in all existing systems in your network, while you are able not to allow an external system to do this action. If you are really worried, you can conFig your host systems in a way that impede ICMP releases completely.

6-2 - Defense against SYN flood attacks

Small blocks

SYN Cookies

A new defense against SYN flood is SYN Cookies. In SYN Cookies each side of the communication, has its own sequence numbers. In response to a SYN, the attacked system, creates a special sequence number from the communication which is a "cookie" and then forgets everything. In other words, eliminate them from the memory is (Cookies are used uniquely to determine an exchange or negotiation). Cookie contains information about the necessary information communication; therefore, later it can recreate the forgotten information about the communication when the packets come from a healthy communication.

6-3 - Coping with DDOS attacks

How to take care of your servers against sent data attack from infected computers in the internet to prevent company's network from disrupting? Here are some ways to deal with DDOS attacks in which are presented in three sections below: Attack prevention, attack detection and attack response.

6 - 3 - 1 - Attack Prevention

Egress filtering Performs filtering on the external traffic and only allow the packets that have a valid source address to leave the network. The extension of property brings about the reduction of the attacks in which the fake IP address is used. However, there is away to fool the Egress filtering and that is the production of attacking packets that their IP address is faked in the network address range of the source [4].

D-WARD detects the external attacks and stops them through controlling the traffic issued to the target machine. It should be installed in the router of the source which works as a gateway between the network and the rest of the internet. This router is conFigd with a set of authorized local source addresses to run the egress filtering on the traffic issued from the source. Also, the networking and communication flows are always monitored to detect unusual behavior . these methods like Egress filtering can be fooled [4,3].

Ingress filtering filters the incoming traffic with invalid IP addresses of the source. These invalid source addresses can be the internal IP address entering from the external network or it can be any special reserved IP address (for example, 192.168. *. *) .

Ingress filtering is a reasonable way to block fake special IP addresses with complete confidence. , but the range of addresses that can be used by the attackers to counterfeit is still too wide. Therefore, even after removing the attack traffic mentioned above, this method is unable to prevent the DDOS attacks effectively.

6 - 3 - 2 Attack Detection

MULTOPS is used to detect bandwidth attacks, in which non-adaptive protocols such as UDP And ICMP. But, in detecting attacks in which a consensus protocol like TCP is used it fails [2].

MULTOPS has three main assumptions which are as follows:

1. attacker and target are separated at least by a router.
2. The rate of the packets is symmetric between two hosts. Meaning that the rate of the packets from A To B is Equal to the rate of packets from B To A. however, the traffic in both directions may not always be equal, like in downloading files or in video .
3. Finding location through a router equipped with MULTOPS is symmetric and constant. It means that if a package comes to B from A passes the router R, packets come to A from B will pass he router R.

6 - 3 - 3 response to the attack

this section discusses the various mechanisms to respond to DDOS attacks.

6 - 3 - 3 -1 Traceback

Each IP packet has two addresses: the source and destination addresses. Destination address is used in route finding in order to deliver the packet to the destination. The route finding infrastructure of IP network does not check the validation of the source address which is placed in the IP packet. The source address is used by the destination machine in order to determine the source for giving answer. In general, no entity is responsible for the source address accuracy. Its scenario is similar to sending mail using mail service. This property is used by the attacker to hide their source address and identity by forging the source IP address. The reason for recommending Traceback mechanisms is to realize the attacker source correctly, provide the possibility of answering, and stop the attack at the nearest point to its source.

6 -3-3-2 Reconfiguration

Reconfiguration mechanisms change the topology of the target or intermediate network to hide the legitimate paths toward the target l from the attacker or isolate the attacker's machine. Such a plan is based on the secure covering service architecture which is used to protect the specified targets from DDOS attacks.

The entry points of covering network and the access point of secure cover (SOAP) perform the identity recognition and allow only legitimate traffic to enter into the network. SOAPs try to find the Beacon to send traffic to them. The Beacons then work confidentially with the Servlet to send traffic to it. Beacons and Servlets of the network remain hidden from the reporters. The specified targets are protected confidentially by means of the filters with high efficiency. They do this through eliminating the traffic.

Randomness and anonymity in this way makes targeting the nodes along the path to a special destination that is protected by SOS difficult for the attacker. Path redundancy is presented in order to hide the identity of confidential Beacons and Servlets. SOS disadvantage is that it requires setting up a covering network and complex algorithms such as: route finding algorithm Chord and Hashing adaptive for finding and assigning Beacons and Servlets. Beacons and Servlets can also be attacked [4].

6 -3-3 - 3 Redirection

Black hole filtering allows the administrator to lead up the attack traffic to a null IP address to remove it. When an attack is detected, a static route is created to lead attack traffic into a "black hole" instead of the victim machine The problem here is that with the appearance of false positive, legitimate traffic will be also discarded like attack traffic.

6 -3 - 3 - 4 Filtering

Filtering mechanisms filter the attack streams completely. Filtering mechanisms rely heavily on third-party detection tools. The filtering function should be done only when the detection result is reliable. Detection can be divided into two main categories: "unorthodox or unconventional behavior-based techniques" and "model-based techniques." Unconventional behavior-based techniques assume that a profile with normal activity is created for the system. Activities that do not match the profile are considered as intruder. However, if an action which is not intrusive but not registered in the normal profile is treated as an attack can lead to false positives. Then filter obstructs the service by its own defense systems. When an intrusive activity but not anomaly occur and gives rise to the attacks that are not detected a false negative appears. In the second technique, the attacks are presented in the form of model. In a way that even similar attacks can be detected. But it can only detect known attacks

and respond to them. For new attacks that the properties of the packets and attack pattern are unknown, it is less used.
However, the pattern-based designs when the traffic matches with the known attack patterns are very useful tools for filtering as a response mechanism [4].
Another solution is to use a firewall to filter out attack traffic. Before entering or leaving the network, packets wait to be processed in accordance with the standards of protection and firewall security.

6 -3 - 3-5 Legitimacy testing

In NetBouncer, a large list of applicants who have been proven to be legitimate is kept. If a packet is received from a source that is not in the legitimate list the types of tests are done to prove the legitimacy of the source. If a source passes these tests successfully, that will be added to the legitimate list and subsequent packets originating from this source are accepted until it the window of legitimacy expires. When it was accepted, the legitimate packets transmission is controlled by a traffic management subsystem to make sure that legitimate applicants are not abusing the consumption of bandwidth and the target does not suffer a traffic that seems to be legitimate. In this way, NetBouncer is able to distinguish legitimate traffic from illegitimate so that it can discard the illegitimate traffic. Tests of legitimacy due to the additional resources that will be allocated for testing give rise to delays in traffic processing and make it slow. [4]

6 -3 -3-6 Attackers' resource consumption

Client puzzles introduce an interactional action based on a cryptographic against connection depletion attacks. Connection depletion is a DOS attack in which the attacker tries to make a lot of faulty communication with the server in order to deplete the resources and disabling them to provide the service to the legitimate requests. The basic idea is that when a server is under attack, that server distributes some little hidden puzzles for users who have requested a service. To complete his application, the user must correctly solve his puzzle. The advantage of this plan is that legitimate traffic can for sure be distinguished from attack traffic. However, like NetBouncer, solving such puzzles requires processing the resources during the attack and causes the system to become slow [4].

7 - **Conclusion**

in this article a series of very Common and in use attacks DDOS and Dos have been explained I , Denial of service attacks is an important and complex issue and thus several techniques have been proposed to deal with them. As the mechanisms to deal with attacks expands, hacker motivation to use these tools will change and probably includes blind transfer of excessive biased competition or defrauding. Without any attention to their reasons, the hackers want to disable the target system. And they try the ways such as stopping the services and complete burst to make the data one-sided. In this paper the methods to handle the attacks were divided into three different groups: attack prevention, attack detection and coping with attack. If a damage can create one-sided streams of DOS information through DDOS attack we Should defend our main system against these attacks. We mentioned some ways to cope with them.

Acknowledgment

I wish to thank and express my deep appreciation to Mohammad Naghizadeh who has, as usual, given me his support and thoughts during the writing of this paper. Moreover, this paper wouldn't be conducted in English if he didn't devote his time to translate my writings into English.

References

[1] FHS Underground Group 2005-2006
Attacks, " *IEEE*
[2] Thomer M. Gil, "MULTOPS: a data structure for denial-of-service attack detection", Ph.D. Thesis, Vrije University, Dec 2,000.
[3] Jelena Mirkovic, "D-WARD: Source-End Defense Against Distributed Denial-of-Service Attacks", Ph.D. Thesis, University of California, Los Angeles, 2003.
[4] Vrizlynn Thing Ling Ling, "Adaptive Response System for Distributed Denial-of-Service Attacks", Ph.D. Thesis, College London, Aug 2008.
[5] Jelena Mirkovic, Janice Martin and Peter Reiher, "A Taxonomy of DDOSAttacks and DDOSDefense Mechanisms", Computer Science Department, University of California, the 2,002th
[6] Christos Douligeris, Aikaterini Mitrokotsa, "DDOSattacks and defense mechanisms: classification and state-of-the-art", 13 October two
thousand and three, Available from:
Http://Www.sciencedirect.com.
C. Joshi, and Manoj Misra, Member, IEEE,
[7] Karthikeyan. KR and A. Indra, "Intrusion Detection Tools and Techniques-A Survey", International Journal of Computer Theory and Engineering, Vol.2, No.6, December 2,010 .
[8] Abraham Yaar, Adrian Perrig, Dawn Song, "StackPi: New Packet Marking and Filtering Mechanisms for DDOSand IP Spoofing Defense", IEEE Journal, Carnegie Mellon University, Vol. 24, Oct 2 006
[9] Jelena Mirkovic and Peter Reiher, "A Taxonomy of DDOSAttack and DDOSDefense Mechanisms", Funded by DARPA, University of Delaware and University of California, 2 004 ..

Energy Efficient Cooperative Caching in Wireless Multimedia Sensor Networks

Narottam Chand

***Abstract*—**Rapid advances in low power hardware technologies, wireless communication and multimedia devices such as microcamera and microphone has made the development of wireless multimedia sensor networks (WMSNs) a reality. Sensor nodes in WMSN capture multimedia data such as image, audio and video, and generate large volume of data for communication and processing in contrast to traditional wireless sensor networks (WSNs). Because of limited battery energy in each sensor node, energy efficiency problem is crucial to be considered in WMSN. In this paper we cope with the energy efficiency problem through cooperative caching of voluminous multimedia data among sensor nodes thus prolonging the network lifetime. The proposed technique uses entropy to determine the importance of multimedia data. Cache admission control and data value based replacement policies ensure the energy efficiency for data communication and processing. Simulation results show that our proposed technique outperforms the existing cooperative caching techniques in terms of various performance metrics.

***Index Terms*—**Multimedia data, cooperative caching, cache replacement, admission control, WSN, WMSN.

I. Introduction

The technological advancements in low power hardware design and wireless communication have enabled the development of tiny, low cost and low power sensor nodes which have capability to sense and compute physical parameters, and are able to communicate with each other. A wireless sensor network (WSN) consists of large number of sensor nodes where node is equipped with constrained on-board processing, storage and radio capabilities. The sensor networks have wide variety of functionalities such as monitoring temperature, pressure, light intensity, movement of object, etc. Such networks may be deployed in unattended harsh environment where it is difficult to recharge or replace batteries of sensor nodes. Due to battery constraints, the design of protocols and applications for such networks have to be energy efficient in order to prolong the lifetime of the network.

Recently, the production of cheap complementary metal oxide semiconductor (CMOS) cameras and microphones has given boost to design of sensor nodes which are capable to capture multimedia content such as image, audio and video [11]. The WSNs where sensor nodes are deployed with a microcamera or microphone to track or monitor any activity using multimedia content, are called wireless multimedia sensor networks (WMSNs) [5]. Compared with traditional WSNs, WMSNs can capture surrounding environment using multimedia information and enable new applications such as multimedia surveillance, advanced health care activities, industrial process control, mobile target tracking, and so on [13]. It is crucial for the majority of applications to serve the requested data with short latency and minimum energy consumption [14]. The success of such applications depends on optimization of the sensor node resources. The cooperative data caching can be used as effective and efficient technique to achieve such goals. Since the battery lifetime can be extended if we reduce the volume of data communication, therefore caching the useful data for each sensor node either in its local storage or in neighboring nodes can prolong the network lifetime.

Caching is a potential technique being employed in traditional areas such as operating systems, virtual memory, distributed systems and Web environments to enhance the system performance by improving the data availability and query response time. Data requests can be served faster from the cache rather than sending a request to the original source, which may be located at larger distance. Cooperative caching is a technique where a group of caches at different nodes work in coordination to achieve better performance in terms of query latency and consumption of computing as well as communication resources.

During recent past, a lot of research in data routing [2], data compression [3] and in-network aggregation [4] has been carried out in WSNs. This paper targets the problem of efficient dissemination of multimedia data and tries to solve it by utilizing the memory of sensor nodes to cache important image data. Such cooperative caching in WMSN can reduce network traffic and enhance data availability to the users through sink.

The traditional cooperative caching techniques used for scalar data such as temperature, pressure, etc. cannot be used for multimedia data such as image, audio and video. Motivated by the unique requirement of multimedia data in WMSN, we propose Entropy based Energy Efficient Cooperative Caching (E3C2) technique for WMSN where data importance is determined by the amount of information contained in an image. Entropy is a qualitative measure of information contained in an image thus is suitable for WMSN.

In the proposed technique, the sensor field is divided into equal size grids called clusters where each cluster is monitored and controlled by a node called cluster head (CH). The CH also maintains cache index within a cluster. Since a CH has to

Narottam Chand is Associate Professor with the Department of Computer Science & Engineering, National Institute of Technology, Hamirpur, 177 005 India. E-mail: nar.chand@gmail.com

perform more data communication and processing compared to a simple sensor node, therefore its energy depletes at faster rate. The role of the CH, therefore, is changed whenever its energy falls below threshold value and can be assigned to other node within a cluster which is richer in energy resource.

Rest of the paper is organized as follows. Section II describes the related work. System model has been explained in Section III. Entropy as a measure of content importance in images has been described in Section IV. Section V describes proposed E3C2 scheme. Section VI defines various simulation parameters, performance metrics and explains simulation results. Finally, Section VII concludes the paper.

II. RELATED WORK

Due to the advances of CMOS technology, microcamera and microphone could be equipped on a sensor node. In addition to scalar quantities, such sensors can collect multimedia data from the environment. Because of large volume of multimedia data, the WMSN node consumes more energy for data transmission and thus energy efficiency is a hot research issue in WMSNs. Cooperative caching of multimedia data in WMSN can reduce the number of transmissions thus can ensure better energy efficiency of nodes. However, the traditional cooperative caching techniques used for scalar data in WSN cannot be applied in WMSN.

In the past decades, a lot of research has been carried out on caching techniques in various fields such as databases, Web applications, operating systems and wireless networks. In WSN, researches have been carried out by exploiting data caching either in some intermediate nodes or at a location nearer to the sink. Jinbao Li et al. [6] proposed a caching scheme for the multi-sink sensor network. The sensor network forms a network tree for particular sink. A common subtree is formed out of such trees and the root of the common subtree is selected as the data caching node to reduce the communication cost.

Md. A. Rahman et al. [7] proposed effective caching by data negotiation between base station and the sensors, developing expectancy of data change and data vanishing. J. Xu et al. [5] proposed a waiting cache scheme which waits for the data of same cluster until it becomes available within a threshold, aggregating it with the packet from the lower cluster and then sending it to the sink, thus reducing number of packets travelling in the network. K.S. Prabh et al. [8] consider the whole network to be a Steiner Data Caching Tree which actually is a binary tree and buffers data at some intermediate node (data cache) such that it reduces the network traffic during multicast. In [9], M.N. Al-Ameen et al. exploit caching for faulty nodes in WSNs and propose a mechanism to handle the packets when node fails. T.P. Sharma et al. [4] proposed a cooperative caching scheme which exploits cooperation among various sensor nodes in a defined region. Apart from its own local storage, a node utilizes memory of nodes from certain region around it to form larger cache storage known as cumulative cache. A token based cache admission control scheme is devised where node holding the token can cache or replace data item. Disadvantage of proposed model is that, there are overheads to maintain and rotate the token. A node importance (NI) based cooperative caching method named NICoCa was proposed in [10][11]. By incorporating the node importance of the WMSN and the residual energy of each sensor node, NICoCa can prolong the network lifetime and reduce query latency. However, only taking the attributes of multimedia item such as data size and the timestamp of the latest access into consideration for cache replacement is insufficient. This is because the multimedia item with larger size has a higher priority to be selected as the candidate victim for replacement. Therefore, importance of data item in terms of content is also equally important while making the cache replacement decisions. In NICoCa, overhead to find NI for all the nodes consumes energy which in turn reduces the lifetime of sensor network.

In our previous work, we have proposed a cooperative caching technique C3S for sensor networks [18]. However, this caching strategy considers only the scalar data generated by the sensor nodes and hence cannot be used directly for multimedia data in WMSNs.

III. SYSTEM MODEL

We assume a wireless multimedia sensor network (WMSN) consisting of sensor nodes (SNs) fitted with micocameras to capture image data from the physical environment. A SN that captures the original image is called source for that particular multimedia data. A data request initiated by a sink is forwarded hop-by-hop along the routing path until it reaches the source and then the source sends back the requested data. Sensor nodes frequently access the data, and cache some data locally to reduce network traffic and data access delay. As sensor nodes do not have sufficient cache storage to store multimedia data over a period of time, cooperative caching may be more useful where cached data at sensor node may also be shared by the neighboring nodes.

WMSN comprises a group of sensor nodes communicating through omni-directional antennas with the same transmission range. The WMSN topology is thus abstracted as an undirected graph G = (V, E), where V is the set of sensor nodes SN_1, SN_2, ..., and $E \subseteq V \times V$ is the set of radio links between the nodes. The existence of a link $e = (SN_i, SN_j) \in E$ also means $(SN_j, SN_i) \in E$, and that nodes SN_i and SN_j are within the transmission range of each other. Here, SN_j is one hop neighbor to SN_i and vice versa. The set of one hop neighbors of a node SN_i belonging to the same cluster/grid is denoted by SN_i^1 and forms a cooperative cache region. The combination of nodes and transitive closure of their cluster neighbors forms a WMSN. A path from $SN_i \in V$ to $SN_j \in V$ has the common meaning of an alternating sequence of vertices and edges, beginning with SN_i and terminating with SN_j. In the proposed model, the CH nodes that lie on the path from source to sink participate during the data communication. A node may take decision regarding caching of data before forwarding it to the next hop. To conserve the energy, sensor nodes within a cluster may be turned off/on at any time, so the set of live nodes varies with time.

IV. ENTROPY OF IMAGE DATA

Entropy is a measure of information in a system and its concept has been employed in many scientific and engineering fields such as communication, image compression, etc. [15][16]. In information theory, the concept of entropy is used to quantify the amount of information necessary to describe the macrostate of a system [15]. If a system presents a high value of entropy, it means that much information is necessary to describe its states.

Multimedia data i.e. images are composed of a set of pixels whose values encode different colors or gray levels. The entropy of a discrete random distribution p(x) is defined as [15] $H(p) = -\sum_{x} p(x)\log_{2} p(x)$.

In terms of multimedia data i.e. image, p(x) can refer to the distribution of gray levels or to the intensity of different color components of an image. The histograms p(x) of a colored image are obtained by counting the number of pixels with a given color intensity (red (R), green (G) or blue (B)), varying from 0 to 255. This procedure generates a set of three different histograms $\{h_c(x), c \in \{R, G, B\}\}$. The entropy may provide a good level of information to describe a given image. If all pixels in an image have the same gray level or the same intensity of color components, this image will present the minimal entropy value. On the other hand, when each pixel of an image presents a specific gray level or color intensity, this image will exhibit maximum entropy. Thus, entropy can be used to characterize the information contained within an image i.e. multimedia data in WMSN. Application to caching is justified because we are interested to cache more important image i.e. one having more information.

V. PROPOSED COOPERATIVE CACHING

We have proposed Entropy based Energy Efficient Cooperative Caching (E3C2) technique that determines the data importance by using entropy value. E3C2 exploits cooperation among various sensor nodes inside a cluster. The design rationale of E3C2 is that, for a sensor node, all other nodes within its cluster domain form a cooperative cache system for the sensor node since local caches of the nodes virtually form a cumulative cache. Each cluster head (CH) acts as the Cache Index (CI), which records the information about cached items by all the nodes within the cluster. When a node in any cluster stores/deletes some data item into/from its cache, it sends the information to its CI so that the corresponding index value can be updated. For each cached item its Time To Live (TTL) information is also maintained at the CI. Whenever, cluster head is rotated, the responsibility of CI is transferred to new cluster head.

When a sensor node issues a request for a data item, it searches its local cache. If the item is found there (a local cache hit), the query response is generated. Otherwise (a local cache miss), the node will look up the required data item from the cluster members by sending a request to the CH. The CH searches in the CI and responds back accordingly. Only when the node cannot find the data item in the cluster members' caches (called cluster cache miss), it will request the data with the CI that lies on the routing path towards source. If a cluster along the path to the source has the requested data (called remote cache hit), then it can serve the request without forwarding it further towards the source. Otherwise, the request will be satisfied by the source.

For a data request, Fig. 1 shows the behavior of E3C2 caching strategy. For each request, one of the following four cases holds:

Local hit: At the first step, the node will check whether the requested data item exists in the cache. If the match is found and stored item is valid, then item is returned to serve the query and process terminates.

Cluster hit: If the match is not found or the stored data item is invalid, the node will ask the CH to confirm whether the requested data item is cached by any other node within the cluster. The CH returns the address of the node that has cached the data item after searching in CI. The requester node will confirm the request and gets data from the caching node.

Remote hit: If the data is not found within the cluster, the request has to flow towards the data source. When the data is found with a node belonging to a cluster (other than home cluster of the requester) along the routing path to the data source, the data is returned to the original requester through the reverse path.

Global hit: When data request is not satisfied, it finally reaches the data source and in response, the data source sends back the data to the requester.

A. Cache Admission Control

Cache admission control decides whether an incoming data item should be stored into the cache of the node or not. Inserting a data item into cache might not always be favorable because incorrect decision can lower the probability of cache hits and also makes poor utilization of the limited storage. In E3C2, the cache admission decision at a node SN_i is based on importance of the data item and distance (number of hops) from where the data item is retrieved.

B. Cache Consistency

Cache consistency is used to confirm whether a data item is valid or not. Due to multi hop environment, strong consistency model cannot be used in WMSN. The E3C2 caching uses a simple weak consistency model based on Time To Live (TTL), in which a SN considers a cached copy up-to-date if its TTL has not expired.

C. Cache Replacement Policy

A cache replacement policy decides the data item that should be deleted from the cache when the cache does not have enough free space to accommodate an incoming item. Such policies apply a value function to each of the cached items, and select as victims, those items which satisfy some criteria. The traditional cache replacement approaches used for scalar data in WSN cannot be used for multimedia data in WMSN [17]. We have developed value based cache replacement policy, where data item with the lowest value is removed from the cache.

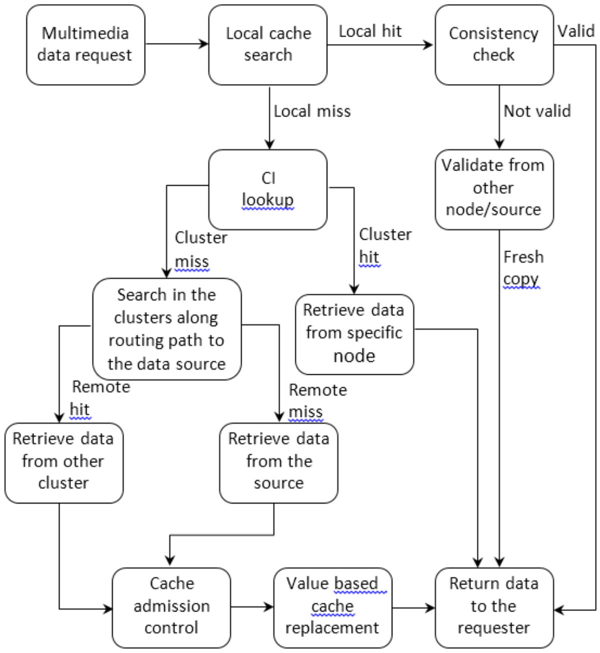

Multimedia data request
Local cache search
Local hit
Consistency check
Valid
Local miss
Not valid
CI lookup
Validate from other node/source
Cluster miss
Cluster hit
Fresh copy
Search in the clusters along routing path to the data source
Retrieve data from specific node
Remote hit
Remote miss
Retrieve data from other cluster
Retrieve data from the source
Cache admission control
Value based cache replacement
Return data to the requester

Fig. 1 Working of E3C2 cooperative caching strategy.

Four factors are considered while computing value of a data item at a node:

Importance. The entropy gives a measure of the importance of a data item. An item d_i with lower entropy H_i should be preferred for replacement.

Popularity. The access probability reflects the popularity of a data item for a node. An item with lower access probability should be chosen for replacement. At a node, the access probability P_i for data item d_i is given as

$$P_i = a_i / \sum_{k=1}^{N} a_k$$

Where a_i is the mean access rate to data item d_i.

Distance. Distance (δ) is measured as the number of hops between the requesting node and the responding node (data source or cache). This policy incorporates the distance as an important parameter in selecting a victim for replacement. The greater the distance, the greater is the importance of the data item. This is because caching data items which are farther away, saves bandwidth and reduces latency for subsequent requests.

Consistency. A data item d_i is valid for a limited lifetime, which is known using the TTL_i field. An item which is valid for shorter period should be preferred for replacement.

Based on the above four factors, the $value_i$ for a data item d_i is computed using the following expression

$$value_i = H_i P_i \delta_i TTL_i$$

The objective is to maximize the total value for the data items kept in the cache. Therefore remove the cached data item d_i having minimum $value_i$ until the free cache space is sufficient to accommodate the incoming data.

VI. PERFORMANCE EVALUAION

In this section, we evaluate the performance of proposed E3C2 protocol through simulation experiments. We compare the performance of E3C2 protocol with NICoCa [10] and C3S [18] which are cooperative caching protocols for wireless sensor networks. We conducted a large number of experiments with various parameters.

A. Simulation Parameters

In our simulations, 400 sensor nodes are randomly deployed in a square area of size 100×100 m^2. Sensor nodes are stationary and have same initial energy 2 J. All simulations are based on a collision free MAC protocol without data loss. Various simulation parameters are listed in Table I.

TABLE I
SIMULATION PARAMETERS

Parameter	Default Value	Range
Network size	(100×100) m^2	(50×50)~(400×400) m^2
Number of nodes	400	100~500
Transmission range (r)	100 m	20~140 m
Sink location	(0, 0)	
Initial energy of node	2 Joule	
Multimedia data size	16 KB	
Mean query generate time (T_q)	5 sec	2~100 sec
Cache size (C)	800 KB	200~1400 KB
TTL	300 sec	100~300 sec
Skewness parameter (θ)	0.8	0~1
E_{elect}	50 nJ/bit	
ε_{fs}	10 pJ/bit/m^2	
ε_{amp}	0.00134 pJ/bit/m^4	

B. Performance Metrics

Network Lifetime

Network lifetime of wireless sensor network is the time span from the deployment to the instant the network works and is able to achieve its objectives. During our simulation, we have used HND (number of rounds after which 50% nodes die) parameter to measure network lifetime.

Average Query Latency (T_a)

The query latency is the time elapsed between the query is sent and the data is transmitted back to the sink, and average query latency (T_a) is the query latency averaged over all the queries.

Byte Hit Ratio (B)

Byte hit ratio is defined as the ratio of the number of data bytes retrieved from the cache to the total number of requested data bytes. It is used as a measure of the efficiency of the cache management. Here byte hit ratio (B) includes local byte hit (B_{local}), cluster byte hit ($B_{cluster}$) and remote byte hit (B_{remote}).

Energy Consumption

It is energy consumption to serve user query by the sink node.

C. Results

Here we study the effect of cache size on various performance metrics on our proposed protocol E3C2 and compare the results with existing cooperative caching protocol NICoCa [10] and C3S [18] for WSN.

Effect of Cache Size on Network Lifetime

To study the effect of cache size on the network lifetime, the number of nodes and the network size is kept fixed at its default values. Fig. 2 shows the result of network lifetime as a function of cache size of sensor nodes. All the schemes exhibit better network lifetime with increasing cache size. This is because more required data items can be found in the local cache with increasing cache size. Due to cooperation within a cluster, the remote byte hit ratio of E3C2 and C3S increases with increasing cache size because each node shares caches of its neighbors within the cluster. When the cache size is small, the contribution due to cluster hit and remote hit is more significant. Due to increase in local, cluster and remote byte hit ratio with increasing cache size, the overall byte hit ratio increases in the proposed protocol E3C2 scheme. As byte hit ratio increases, more data may be shared from the nearby sensor nodes, thus reducing the number of transmissions and hence prolonging the network lifetime.

E3C2 always performs better than NICoCa and C3S due to consideration of entropy to cache more significant data. NICoCa behaves worst due to the fact that to compute node importance (NI), the message overhead is large and energy of a node depletes in exchanging these messages.

Effect of Cache Size on Average Query Latency

The effect of cache size on average query latency for proposed protocol E3C2, and existing protocols NICoCa [10] and C3S [18] has been shown in Fig. 3. With increasing cache size more number of requests are satisfied from the cache thus decreasing the average query latency. Due to the use of more suitable cache replacement, E3C2 behaves better than other strategies.

Effect of Cache Size on Byte Hit Ratio

Fig. 4 shows effect of cache size on byte hit ratio. All the schemes exhibit better byte hit ratio with increasing cache size. Due to cumulative caching within a cluster and better replacement policy, the E3C2 scheme outperforms other schemes under different cache size settings. We deploy value based replacement in E3C2 which retains more useful data in the caches of nodes and thus increasing the overall byte hit ratio.

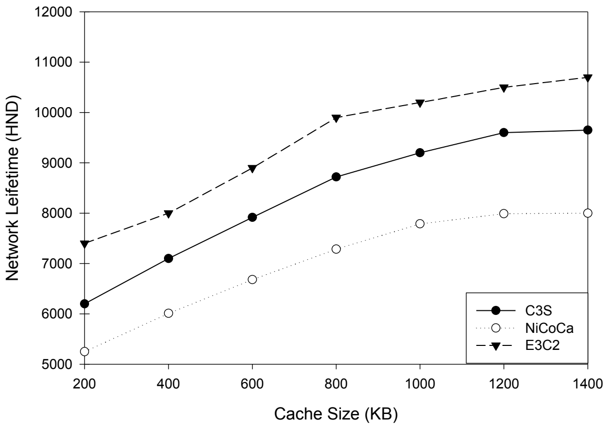


Fig. 2 Effect of cache size on HND.

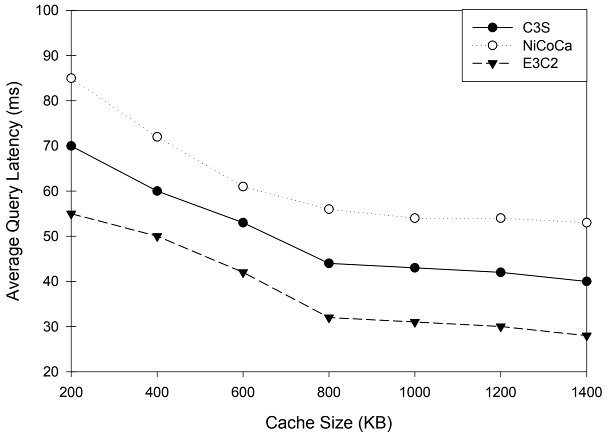


Fig. 3 Effect of cache size on average query latency.

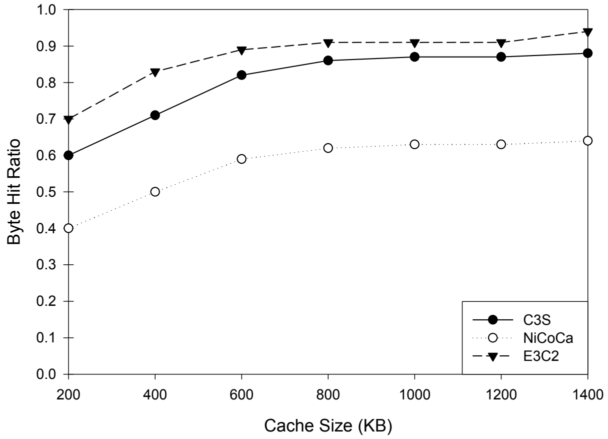


Fig. 4 Effect of cache size on byte hit ratio.

Effect of Cache Size on Energy Consumption

Fig. 5 shows effect of cache size on energy consumption. At smaller cache size, less number of requests are satisfied from the local cache, thus increasing the energy consumption. As byte hit ratio increases, more data may be shared from the nearby sensor nodes, thus reducing the number of transmissions and hence energy consumption.

E3C2 and C3S have always lower energy consumption than

NICoCa because message overhead is very high in NICoCa to compute node importance (NI) which consumes more communication energy. Due to cumulative caching within a cluster and better replacement policy, the E3C2 scheme outperforms NICoCa scheme and C3S scheme under different cache size settings.

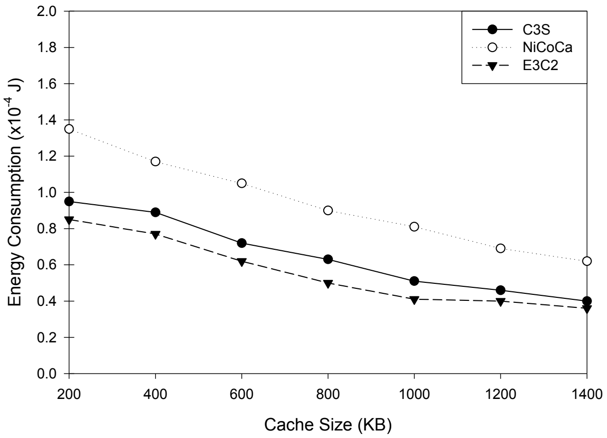


Fig. 5 Effect of cache size on energy consumption.

VII. CONCLUSION

For the energy constrained WMSNs, the most challenging problem is to effectively use the energy of network during data collection and query processing. In this paper, we have proposed a cooperative caching scheme E3C2 for supporting efficient data collection and query processing in WMSNs. The scheme enables sensor nodes to cooperatively share multimedia data thus reducing the query latency and energy consumption of the nodes. We have used entropy of images to determine their importance and take caching decision. It has been observed that lifetime of WMSN is enhanced through cooperative caching. The proposed cache discovery algorithm ensures that a requested data is returned from the nearest cache or source. The admission control prevents high data replication by enforcing a minimum distance between the same data item, while the replacement policy helps in improving the byte hit ratio and accessibility. Cache consistency ensures that nodes only access valid states of the data. Simulation results show that the E3C2 caching scheme performs better in terms of various performance metrics in comparison with NICoCa and C3S strategies.

REFERENCES

[1] J. Chen and H. Zhou, "Cooperative Energy Efficient Management Scheme for Multimedia Information Dissemination," International Journal of Distributed Sensor Networks, 10 pages, 2014.

[2] Abbasi and M. Younis, "A Survey on Clustering Algorithms for Wireless Sensor Networks," ACM Journal of Computer Communications, Vol. 30, No. 14-15, pp. 2826-2841, 2007.

[3] N. Kimura and S. Latifi, "A Survey on Data Compression in Wireless Sensor Networks," International Conference on Information Technology: Coding and Computing, Vol. 2, pp. 8-13, 2005.

[4] T.P. Sharma, R.C. Joshi and M. Misra, "Dual Radio Based Cooperative Caching for Wireless Sensor Networks," IEEE International Conference on Networking, pp. 1-7, 2008.

[5] J. Xu, K. Li, Y. Shen and J. Liu, "An Energy-Efficient Waiting Caching Algorithm in Wireless Sensor Network," International Conference on Embedded and Ubiquitous Computing, Vol. 1, pp. 323-329, 2008.

[6] J. Li, S. Li and J. Zhu, "Data Caching Based Queries in Multi-Sink Sensor Networks," International Conference on Mobile Ad-hoc and Sensor Networks, pp. 9-16, 2009.

[7] Md. A. Rahman and S. Hussain, "Effective Caching in Wireless Sensor Network," International Conference on Advanced Information Networking and Applications Workshops, Vol. 1, pp. 43-47, 2007.

[8] K. Prabh and T. Abdelzaher, "Energy-Conserving Data Cache Placement in Sensor Networks," ACM Transactions on Sensor Networks, Vol. 1, No. 2, pp. 178–203, 2005.

[9] M.N. Al-Ameen and Md. R. Hasan, "The Mechanisms to Decide on Caching a Packet on Its Way of Transmission to a Faulty Node in Wireless Sensor Networks Based on the Analytical Models and Mathematical Evaluations," International Conference on Sensing Technology, pp. 336-341, 2008.

[10] N. Dimokas, D. Katsaros, L. Tassiulas and Y. Manolopoulos, "High Performance, Low Complexity Cooperative Caching for Wireless Sensor Networks," Springer International Journal of Wireless Networks, Vol. 17, No. 3, pp. 717-737, 2011.

[11] N. Dimokas, D. Katsaros and Y. Manolopoulos,"Cooperative Caching in Wireless Multimedia Sensor Networks," Springer Journal of Mobile Network Applications, Vol. 13, No. 3/4, pp. 337-356, 2008.

[12] Xiao, H. Chen and S. Zhou, "Distributed Localization Using a Moving Beacon in Wireless Sensor Networks", IEEE Transactions on Parallel and Distributed Systems, Vol. 19, No. 5, pp. 587-600, 2008.

[13] K Lin and M. Chen, "Research on Energy Efficient Fusion-driven Routing in Wireless Multimedia Sensor Networks," Journal of Wireless Communication and Networking, No. 1, pp. 1-12, 2011.

[14] N. Dimokas and D. Katsaros,"Detecting Energy-Efficient Central Nodes for Cooperative Caching in Wireless Sensor Networks," IEEE International Conference on Advanced Information Networking and Applications (AINA), pp. 484-491, 2013.

[15] AL Barbieri, GF de Arruda, FA Rodrigues, OM Bruno and L da F Costa, "An Entropy-based Approach to Automatic Image Segmentation of Satellite Images," Physica A, pp. 512-518, 2011.

[16] LW Leung, B. King and V. Vohra, "Comparison of Image Data Fusion Techniques Using Entropy and INI," Asian Conference on Remote Sensing, 2011.

[17] VS Tseng, M-H Hsieh and KW Lin, "A novel Cache Replacement Algorithm for Cooperative Caching in Wireless Multimedia Sensor Networks," International Journal of Innovative Computing, Information and Control, Vol. 7, No. 2, pp. 763-776, 2011.

[18] N. Chand, "Energy Efficient Cooperative Caching in WSN," International Conference on Computer and Communication Networks Engineering (ICCCNE), pp. 674-679, 2013.

Dr. Narottam Chand received his Ph.D. degree from IIT Roorkee in Computer Science and Engineering. Previously he received M.Tech. and B.Tech. degrees in Computer Science and Engineering from IIT Delhi and NIT Hamirpur, respectively.

Presently he is working as Associate Professor, Department of Computer Science and Engineering, NIT Hamirpur. He has served as Head, Department of Computer Science & Engineering, during Feb 2008 to Jan 2011 and Head, Institute Computer Centre, during Feb 2008 to July 2009.

His current research areas of interest include mobile computing, mobile ad hoc networks and wireless sensor networks. He has published more than 150 research papers in International/National journals, guided five PhDs and guiding few more in these areas. He is member of ACM, IEEE, ISTE, CSI, International Association of Engineers and Internet Society.

Mathematical Model for Object Oriented Class Cohesion Metric -MCCM

Dr. Omer Saleh, Tejdeda Alhussen Alhadi, Xavier Patrick Kishore, Sagaya Aurelia

***Abstract*—**Despite inherent difficulties, measuring software quality is important because it makes evaluation and improvement of various aspects of software. In this study, a new cohesion metric named MCCM is proposed and its first empirical validations arebeing a structural metric, MCCM can be used in every stage of software lifecycle.

***Keywords*—**MCCM; Metric; Cohesion; Connectivity; Direct; Indirect; Method Invocation; Attribute Usage

I. Introduction

In order to improve the software's quality, the software quality should be measured at various stages of software lifecycle and the obtained measures should be compared and evaluated. One or more metrics should be chosen before making measurement. At this time, one is faced with the difficulty of finding properties that are agreed upon about software quality, which is already hard to define. In order to begin measuring at the earlier phases of software lifecycle, a metric needs to use the internal quality factors. Coupling [1] and cohesion [2] are fundamental properties used for this aim. In the software that are produced by the proper usage of Object Oriented Programming (OOP) approach, low coupling and high cohesion is expected [3]. There are lots of metrics for coupling and cohesion in the literature, but more empirical studies are needed as the verification of software metrics is not easy [4]. In this paper, for the estimation of class cohesion is based on different relationships that may exist between its methods. It takes into account several ways of capturing the functional cohesion of the class, by focusing on Connectivity between methods.

II. Class CohesionMetricDefinition

The degree to which the methods within a class are related to one another and work together to provide well-bounded behavior. Effective object-oriented designs maximize cohesion since it promotes encapsulation. Cohesion is an internal software attribute that depicts how well connected the components of a software module are. This can be determined by knowing the extent to which the individual components of a module are required to perform the same task [5].

Cohesion, in object-oriented terms, is a measure of how strongly related and focused the responsibilities of a module are.

A class with low cohesion does many unrelated things, or does too much work. Such classes are undesirable; they suffer from the following problems:

- Hard to comprehend
- Hard to reuse
- Hard to maintain
- Delicate; constantly effected by change [6].

III. Developed Metric: MCCM

Three steps are followed while calculating the cohesion of a class using the new metric. In the first step, the relationship between methods are examined by looking into the method invocations and the first graphic (G_{xA}) is obtained. In the second step, the relationship between methods are examined by looking into the attribute usage, and the second graphic (G_{xB}) is obtained. In the third and last step, the graphics created in the previous steps are combined into one graphic (G_x). In the remaining sections these steps are elaborated in detail.

A. Examination of Method Invocations

A directed line is created between methods from the calling method to the called method if one of the methods calls the other. The obtained graphic G_{x1} can be expressed asbelow [7]: If a directed graphic G_{x1}(V, E) and V = M_x, E={<m, n> € V × V|(m calls n)}

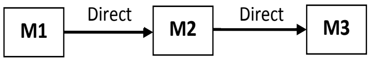
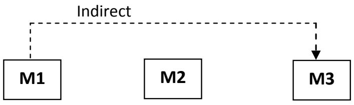


Fig.1. Direct connection graphic (G_{x1})

To find indirect connections, the directed lines are used in graphic G_{x1}. By looking at Fig.2, it can easily be seen that method M1 is calling method M2 and method M2 is calling method M3. Consequently, method M1 is calling method M3 indirectly. Then, a directed line is created from method M1 to method M3. The connection graphic G_{x2} obtained in this way can be expressed as such:If a directed graphic G_{x2}(V, E) and V=Mx,

$$E = \left\{ < m, l > \in V \times V \middle| \begin{matrix} (< m, n > \in G_{x1}) \wedge \\ (< n, l > \in G_{x1}) \end{matrix} \right\} \quad (1)$$

The graphic G_{x2} obtained from graphic G_{x1} can be seen in Fig.2.



Dr. Omer Saleh, Department of Computer Science, Faculty of Education, Beniwalid, Libya (immer.jomah@gmail.com)

Tejdeda Alhussen Alhadi, Department of Computer Science, Faculty of Education, Beniwalid, Libya (tt_hussen@yahoo.com)

Xavier Patrick Kishore, Department of Computer Science, Faculty of Education, Beniwalid, Libya (patrick.kishore@gmail.com)

Sagaya Aurelia1, Department of Computer Science, Faculty of Education, Beniwalid, Libya (Sagaya.aurelia@gmail.com)

Fig.2.Indirect connection graphic (G_{x2})

As the last step of investigation of method invocations, the direct and indirect connected methods are shown by combining the previously obtained graphics G_{x1} and G_{x2} into one graphic G_{xA}. The graphic G_{xA} can be seen in Fig.3 which is obtained from example graphics G_{x1} and G_{x2} [8][9].

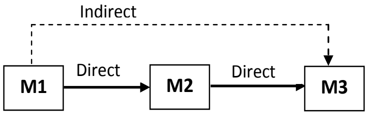


Fig.3. Method invocation graphic (G_{xA})

B. *Examination of Attribute Usage*

A graph showing a sample case of method-attribute relation is drawn in order to see attribute usage by methods. If a method uses an attribute, an undirected line is created between the method and the attribute [9]. An example graphic G_{xB} which is obtained in this way can be seen in Fig.4.

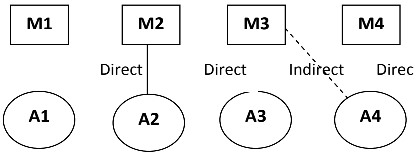


Fig.4 Direct and Indirect attribute usage graphic (G_{xB})

C. *Method Invocation and Attribute Usage*

The graphic G_x is obtained by combining the graphics G_{xA} and G_{xB} created in the previous steps into one graphic. By the examination of indirect method invocations and determination of indirect attribute usage, the final version of G_x is obtained. A G_x graphic which is generated by using the example graphics given in previous steps can be seen in Fig.5 [10][11].

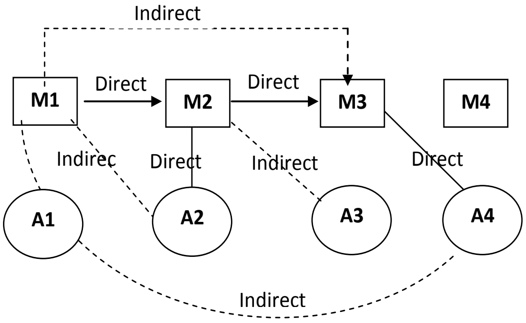


Fig.5. Final relationship graphic (G_x)

IV. Method Connectivity Cohesion Metric (MCCM)

Let X denote a class, I_x the set of attributes, and M_x the set of methods of the class. When a directed graphic $G_x(V, E)$ is considered, the method invocation relation between two methods can be defined as below [11]:

Method_inv(Mi,Mj)=

$$\begin{cases} 1 \text{ , if } < M_i, M_j > \in G_x \quad \text{or} < M_j, M_i > \in G_x \\ \quad 0 \text{ , otherwise} \end{cases} \qquad (2)$$

Let n be number of attributes shared by Mi and Mj .

$$\text{Attr\_usg (Mi , Mj)= n} \qquad (3)$$

The relation between two methods, Relation (Mi, Mj), is the sum of method invocation and attribute sharing between methods Mi and Mj, where Mi ∈ Mx, Mj ∈ Mx, and i ≠ j [12].

$$\text{Relation(Mi,Mj) = Method\_inv(Mi,Mj) + Attr\_usg(Mi, Mj)} \quad (4)$$

The total relation value of the class is found from the relation of two methods as below:

$$\text{Total\_Relation(x)} = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \left[\text{Relation } (M_i, M_j) \right] \qquad (5)$$

$$\text{Total\_Relation(x)} = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \left[\text{Method}_{\text{inv}}(M_i, M_j) + \text{Attr\_usg}(M_i, M_j) \right] \qquad (6)$$

The maximum number of public methods pairs is [13]:

$$n * (n - 1) / 2 \qquad (7)$$

This means an invocation exists in every (Mi, Mj) relation and this can be found by the number of all possible (Mi, Mj) relations:

$$\text{Max (Method\_inv)} = \frac{n.(n-1)}{2} \qquad (8)$$

Similarly, maximum attribute sharing relation in the class can be defined as; Max (Attr_ usg) = means all attributes of the class are shared in every (Mi, Mj) relation:

$$\text{Max\_(Att\_usg)=} \frac{n.(n-1\ \ )}{2} \times a \qquad (9)$$

The value of Max_Relation(X) can be defined as below:

$$\textbf{Max\_Relation (x)} = \left(\frac{n.(n-1)}{2} \right) + \left(\frac{n.(n-1)}{2} \times a \right) (10)$$

According to above explanations

$$\textbf{MCCM} = \frac{\textbf{Total\_Relation}}{\textbf{Max\_Relation}}$$

CI. Theoretical Validation

Several researchers [4][14][15] have proposed properties that software metrics should posses in order to increase their level of confidence; the points considered while developing MCCM are verified.

- *Property 1: Non-negativity*

The cohesion of a class of an object oriented system should belong to a specified interval (i.e. Cohesion (C) ∈ [0, Max]. Normalization allows meaningful comparisons between the cohesions of different classes, since they all belong to the same interval.

- *Property 2: Normalization*

Determines if the result of the metric is normalized i.e. values returned by the metric is between 0 and 1; classes with zero cohesion value have the least cohesion while classes with cohesion value 1 have perfect cohesion.

- *Property 3: Null value and maximum value*

The cohesion of a class of an object oriented system is null if there is no interactions among the components of the class (i.e. interaction among the methods and attributes of the class) and it is maximum if the interaction among the components is maximal.

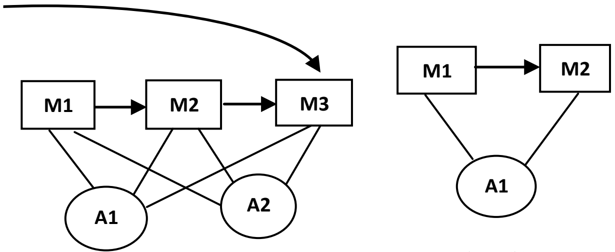


Fig.7.(a)

Fig.7.(b)

Fig.7. Example graphics for maximum relation

When Fig.7 is examined, one intuitively expects the metric to give the maximum cohesion value “1” for both cases a and b. Metric computation for the graphic in Fig.7.(a) (m = 3, a = 2) is as follows:

$$\text{MCCM} = \frac{[(1+2)+(1+2)+(1+2)]}{\left(\frac{3\times(3-1)}{2}\right)+\left(\frac{3\times(3-1)}{2}\times 2\right)} = 1$$

When MCCM is computed for graphic in Fig.7.(b) (m = 2, a = 1), it is calculated as:

$$\text{MCCM} = \frac{[\,(1+1)\,]}{\left(\frac{2\times(2-1)}{2}\right)+\left(\frac{2\times(2-1)}{2}\times 2\right)} = 1$$

As a result, the proposed metric gave the same value, depicting maximum cohesion for both of the graphics in Fig.7, in parallel with the intuition.

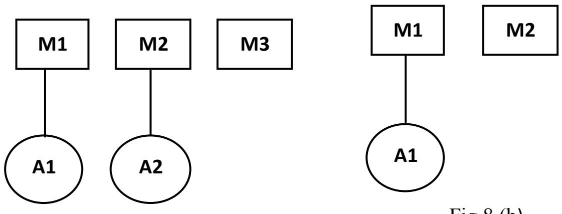


Fig.8.(a)

Fig.8.(b)

Fig.8. Example graphics for minimum relation

In the case of Fig.8, the intuition leads one to expect minimum cohesion as there are no method interactions involved. Metric computation for the graphic in Fig.8.(a) (m = 3, a = 2) is as follows:

$$\text{MCCM} = \frac{[(0+0)+(0+0)+(0+0)]}{\left(\frac{3\times(3-1)}{2}\right)+\left(\frac{3\times(3-1)}{2}\times 2\right)} = 0$$

When MCCM is computed for graphic in Fig.8.(b) (m = 2, a = 1), it is calculated as:

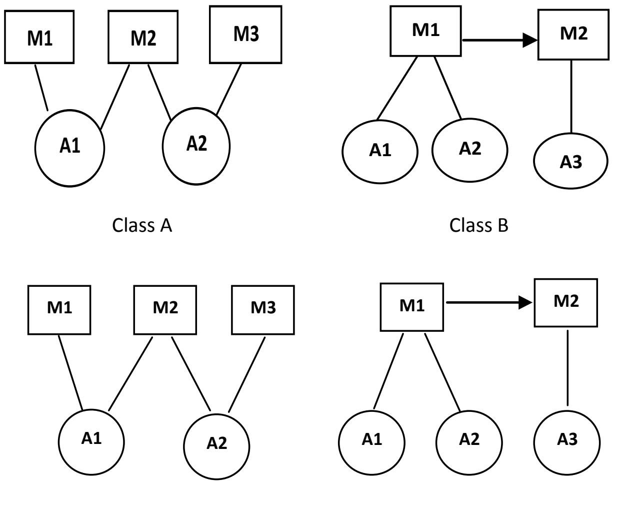
$$\text{MCCM} = \frac{[(0+0)]}{\left(\frac{2\times(2-1)}{2}\right)+\left(\frac{2\times(2-1)}{2}\times 1\right)} = 0$$

As a result, the proposed metric gave the same value, depicting minimum cohesion for both of the graphics in Fig.8, in parallel with the intuition.

The MCCM of a class = 0 if there is no interactions among the components of the class and MCCM = 1 if all methods are directly or indirectly connected (i.e. if the interactions among the components of the class is maximal). Therefore, the value of MCCM lies in the interval [0, 1] inclusive. Hence, the metric satisfies the Property 1, 2 and 3.

- *Property 4: Transitivity*

Consider three classes c1, c2 and c3 such that, Cohesion (c1) <Cohesion (c2) and Cohesion (c2) < Cohesion (c3), then Cohesion (c1) < Cohesion (c3).



Class A

Class B

Class C (The union of Class A and B)

Fig.9. Example graphics for Transitivity

If we have three classes A, B and C, such that Cohesion (Class A) < Cohesion (Class B) and Cohesion (Class B) < Cohesion (Class C). Then it implies that Cohesion (Class A) < Cohesion (Class C). Therefore, MCCM satisfies the property 4.

- *Property 5: Relative number of cohesive interactions*

If the interactions among the components of the class increase, the metric should indicate higher cohesion [16][17].

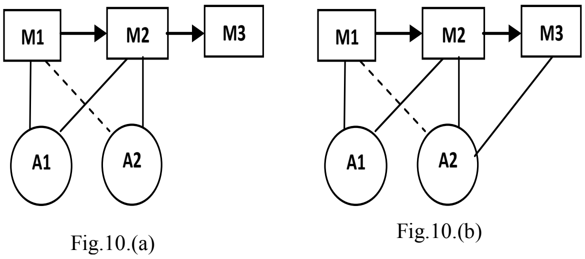


Fig.10. Example graphics for interaction

When Fig.10. is examined, one intuitively expects the metric to give a cohesion value better for Fig.10.(b) than for Fig.10.(a) as there are more interactions in Fig.10.(b) than there are in Fig.10.(a). Metric computation for the graphic in Fig.10.(a) (m = 3, a = 2) is calculated as:

$$\text{MCCM} = \frac{[(1+2) + (1+0) + (0+0)]}{\left(\frac{3\times(3-1)}{2}\right) + \left(\frac{3\times(3-1)}{2}\times 2\right)} = \frac{4}{9}$$

When MCCM is computed for graphic in Fig.10.(b) (m = 3, a = 2), it is calculated as:

$$\text{MCCM} = \frac{[(1+2) + (1+1) + (0+1)]}{\left(\frac{3\times(3-1)}{2}\right) + \left(\frac{3\times(3-1)}{2}\times 2\right)} = \frac{6}{9}$$

As a result, it is seen that MCCM depicts higher cohesion as the interactions among the components of the class increase.

CII. Point of Consideration: Considering indirect relations

If the indirect connections are not taken into account, the cohesiveness cannot be measured precisely [18]. In this case, a metric should indicate lower cohesion. When Fig.11 is examined, one intuitively expects the metric to give the cohesion value better for Fig.11.(b) than for Fig.11.(a). However, if the indirect relations are not taken into account, this expected result cannot be seen.

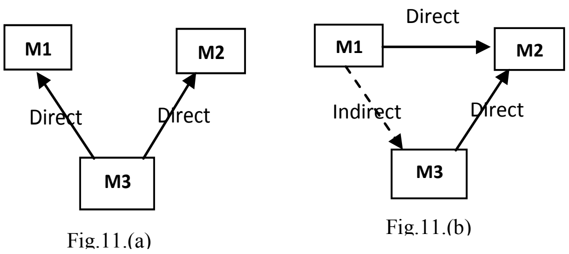


Fig.11. Example for considering indirect relations [19]

If the indirect connections are not taken into account, metric computation for the graphic in Fig.11.(a) is calculated as follows:

$$\text{MCCM} = \frac{[(0+0) + (1+0) + (1+0)]}{\left(\frac{3\times(3-1)}{2}\right) + \left(\frac{3\times(3-1)}{2}\times 0\right)} = \frac{2}{3}$$

If the indirect connections are not taken into account, metric computation for the graphic in Fig.11.(b) is calculated as follows:

$$\text{MCCM} = \frac{[(1+0) + (0+0) + (1+0)]}{\left(\frac{3\times(3-1)}{2}\right) + \left(\frac{3\times(3-1)}{2}\times 0\right)} = \frac{2}{3}$$

When the indirect connections are not taken into account, the metric gave the same result. However, it is desirable for a good metric to differentiate these results. If the indirect connections are taken into account, metric computation for the graphic in Fig.11.(a) stays as the same, as there are no indirect connections in this graphic. If the indirect connections are taken into account, metric computation for the graphic in Fig.11.(b) is calculated as follows:

$$\text{MCCM} = \frac{[(1+0) + (1+0) + (1+0)]}{\left(\frac{3\times(3-1)}{2}\right) + \left(\frac{3\times(3-1)}{2}\times 0\right)} = 1$$

This example shows the importance of considering the indirect connections: Inclusion of indirect relations into MCCM has enabled the metric to produce results in parallel with the intuition.

CIII. CONCLUSION

As computing becomes more pervasive, a software defect can lead to bad results such as financial loss, time delay or the loss of human life in even worse cases. Therefore, software systems should operate as error free and consistent systems. Increasing demand on software quality introduces us the “quality” characteristic as an important factor for a software product. Cohesion, an important quality factor for Object Oriented Programming paradigm, can be described as the degree of connectivity among the elements of a single module. It is expected that cohesion to be high and coupling to be low in a well designed system [9].

The current state of the MCCM metric is encouraging as it gives results that are normalized and parallel with intuitions. MCCM uses both direct and indirect method relations in its calculations; however, it does not take inherited methods into account. For these reasons, continuation of empirical works and the ability to use inherited methods in calculations are planned as future work.

REFERENCES

[1] Al Dallal, J. and Briand, L., A Precise Method-Method Interaction-Based Cohesion Metric for Object-Oriented Classes, ACM Transactions on Software Engineering and Methodology (TOSEM), 2012, Vol. 21, No. 2, pp. 8:1-8:34.

[2] Badri, L. and Badri, M., A Proposal of a New Class Cohesion Criterion: An Empirical Study, Journal of Object Technology, 3(4), 2004, pp. 145-159.

[3] Al Dallal, J., Measuring the Discriminative Power of Object-Oriented Class Cohesion Metrics, IEEE Transactions on Software Engineering, 2011c, Vol. 37, No. 6, pp. 788-804.

[4] Briand, L. C., Daly, J. W., and Wüst, J. K., "A Unified Framework for Coupling Measurement in Object-Oriented Systems." IEEE Transactins on Software Engineering, vol. 25, no. 1, 1999.

[5] Fenton E. N., Pfleeger S. L.: "Software Metrics – A Rigorous & Practical Approach", PWS Publishing company, Boston, 1997.

[6] T. Meyers and D. Binkley, An empirical study of slice-based cohesion and coupling metrics, ACM Transactions on Software Engineering Methodology, 17(1), 2007, pp. 2-27.

[7] Baldassari, B., Robach, C. and du Bosquet, L., "Early metrics for Object Oriented Designs", Proc. 1st Int'l. Workshop on Testability Assessment (IWoTA), 2004, pp. 62-69.

[8] Kramer, S. and Kaindl, H., "Coupling and cohesion metrics for knowledge-based systems using frames and rules", ACM Trans. on Soft. Engineering and Methodology (TOSEM), vol. 13,no. 3, July 2004, pp. 332-358.

[9] Poshyvanyk D., Marcus A., The Conceptual Chesion of Classes. 21st IEEE Int'l. Conf. on Software Maintenance (ICSM'05), 2005, pp. 133-142.

[10] R. Subramanyam and M. Krishnan. Empirical Analysis of CK metrics for Object-Oriented Design Complexity : Implications for Software Defects. IEEE Transactions on Software Engineering, 29(4):297-310, Apr 2003.

[11] Linda Badri and Mourad Badri: "A Proposal of a New Class Cohesion Criterion: An Empirical Study", in Journal of Object Technology, vol. 3, no. 4, April 2004,

[12] Fernndez, L., and Pea, R., A Sensitive Metric of Class Cohesion, International Journal of Information Theories and Applications, 13(1), 2006, pp. 82-91.

[13] Bieman J. M., Kang B., "Cohesion and Reuse in an Object-Oriented System", in Proc. ACM Symp. Software Reusability (SSR'94), 259-262, 1995.

[14] Al Dallal, J., Mathematical Validation of Object-Oriented Class Cohesion Metrics, International Journal of Computers, 2010, Vol. 4, No. 2, pp. 45-52.

[15] Al Dallal, J. Theoretical validation of object oriented lack-of-cohesion metrics, proceedings of the 8th WSEAS International Conference on Software Engineering, Parallel and Distributed Systems (SEPADS 2009), Cambridge, UK, February 2009.

[16] [16] Briand, L. C., Wst, J., and Lounis, H., Replicated Case Studies for Investigating Quality Factors in Object-Oriented Designs, Empirical Software Engineering, 6(1), 2001, pp. 11-58.

[17] Marcus, M., Poshyvanyk, D., and Ferenc, R., Using the Conceptual Cohesion of Classes for Fault Prediction in Object-Oriented Systems, IEEE Transactions on Software Engineering, 34(2), 2008, pp. 287-300.

[18] Al Dallal, J., Incorporating Transitive Relations in Low-Level Design-Based Class Cohesion Measurement, Software: Practice and Experience, 2013a, Vol. 43. No. 6, pp. 685-704.

[19] Counsell, S., Swift, S., and Crampton, J., The Interpretation and Utility of Three Cohesion Metrics for Object-Oriented Design, ACM Transactions on Software Engineering and Methodology(TOSEM),Vol.15,No.2,2006,pp.123-149.

Dr. Omer Saleh Mahmod Jamah (January 25,1973) is now the Director of Post graduate cum Research and Development and Head of the department of Computer science, Faculty of education, Azzaytuna university, Baniwalid, Libya. He received his B.Sc. in Control System and Measurement (1995), M.Sc. in Electrical and Computer Measurement (2004), and Ph.D. in Electrical engineering, Automatics computer science and electronics from AGH University of technology, Krakow, Poland. He has done his Diploma in Planning and time management from Canada Global Centre, Canada. Now he is heading Computer Science department, Faculty of Education, Azzaytuna University, Baniwalid, Libya. His research interest includes multicriteria optimization for solving optimal control problems and Fuzzy logic. He has published 12 papers and attended various national and international Level conferences and workshops.

Ms. Tejdeda Alhussen Alhadi (Feburary 1, 1980) is now with Faculty of Education, Azzaytuna University, Bani-walid, Libya. She is into teaching profession for more than 13 years. She has done B.Sc and M.Sc in Computer Science from Libyan Academy. She has also been involved in various administration related activities. Her specialization includes Database, Software Engineering and Artifical intelligence. She has published various national and international papers and guided many projects.

Mr. Xavier Patrick Kishore (November 6, 1973) received his BSc Mathematics (1994), Master of Computer Application (2002) and Diplomas in E-Commerce and Advanced software Technology. He has received Brain bench certification in Java and HTML. Now he is working in Department of computer science Faculty of Education, Azzaytuna University,Baniwalid, Libya. He is specializedin programming languages. His current research interest includes Natural languageprocessing. He has authored more than 9 papers and attended many conferences.

Er. Mrs. Sagaya Aurelia (November 9,1978) par-time research scholar in Bharathidasan university . Now she is with department of Computer Science, Faculty of Education, Azzaytuna University, Bani-walid, Libya. She received her Diploma in Electronics and Communication (1997),B.E (Bachelor of Engineering specialized in Electronics and Communication Engineering(2000) and M.Tech in Information Technology(2004),she has also done her Postgraduation diplomas in Business Administration (PGDBA) and Journalism and Mass Communication (PGDJMC). She has received Brainbench certification in HTML. Her current research interest includes Virtual reality, Augmented reality and Human Computer Interaction and User interface Design. She has authored14 papers and attendance several national and international level workshops and conferences.

Implementing Hierarchical Access Control in Organizations using Symmetric Polynomials and Tree Based Group Diffie Hellman Scheme

Jeddy Nafeesa Begum, Krishnan Kumar, Vembu Sumathy

***Abstract*—**Implementing Hierarchical Access Control (HAC) is very vital for organizations which follow a hierarchical type of controllability. In organizations , employees are organized into disjoint classes according to their roles and responsibilities. Some Employees are higher in the hierarchy and some are lower. The control methodology for HAC should ensure that resources of employees belonging to lower classes are accessible by employees of higher classes. In the proposed Scheme, the methodology used for implementing the HAC uses a dual encryption scheme comprising of two keys namely a symmetrical polynomial key and a Tree based group Diffie Hellman key. The main component is the presence of a software agent in each class that aids in the functioning. The paper discusses the numerical illustration and implementation of the proposed Scheme. It is found that the proposed scheme can manage the information or resource in an efficient manner and scales well for a large number of employees which are organized under different classes. It shields the resources from the various type of threats and attacks that may try to create havoc in the system. The computation time and communication time for ensuring HAC is also reduced by the proposed Modular Approach

***Keywords*—**. *Hierarchical Access Control , Symmetric Polynomial, TGDH, Software Agent*

I. Introduction

Several aspects such as globalization of markets and production, shift in competition criteria from "cost" to "quality" and "time", progress of human factors as operators and clients, new modern technologies, advances and amalgamation of the enabling technologies of computers and communications, have shaped the modern enterprise during the last decades. A lot of efforts have lately been made by academia, industry people, professional associations and standardization bodies to produce conceptual frameworks, methodologies, commercial support tools and standards for enterprise integration.

In recent times there has been an extraordinary growth in the number of distributed large scale applications based on the Internet for the deployment .Accompanying this growth has been the increased need for distributed access control architectures particularly for corporate networks which are spread across many places and mostly rely on good access control schemes for safe operation. Hierarchical access control prevents the users belonging to lower cadre to access the information of users belonging to higher cadre but allows users belonging to higher positions access to resources of lower class users. This type of scenario is common in many organizations. To implement the HAC , a dual encryption scheme that uses symmetric polynomial keys and Tree based group Diffie- Hellman Schemes are used.

The most important part of the implementation is a software agent that prevents unauthorised users from interacting with particular resources whilst guaranteeing that authorised users will not be denied the access rights. The proposed scheme is found to be fast, efficient and secure.

On approaching the problems of administration and access control of large scale internet systems, there are two main issues to be considered: a) the relevance of hierarchical systems methodology, and b) the ever more increasing use of human proficiency stored in computer programs. This paper aims at an implementation architecture which solves various problems that may occur in a hierarchy , the prominent being the confidentiality of information. The first characteristic of this architecture is based on a splitting up procedure to build a hierarchical structure allowing to manage the complexity of the process. The second characteristic concerns the organization of this structure to achieve HAC. On the one hand a software agent to manage, monitor and to exhibit the HAC capabilities are distributed in each node of this hierarchical structure . So each of these nodes is a " control and monitoring module". On the other hand, a further process model more complete than the one included in each control and monitoring module is put for the hierarchical structure

Jeddy Nafeesa Begum is currently doing her Ph.D in Computer Science , She has about 16 years of teaching Experience . She has about 5 international publications to her credit .

Dr. Krishnan Kumar is an Assistant professor . He completed his doctorate from Anna University , Chennai, Tamil Nadu, India . His research interest are computer networks, cryptography etc., He has about 7 publications in International Journals.

Dr. Vembu Sumathy is presently working as Associate Professor in Government College of Technology, Coimbatore, Tamil Nadu, India.. Ph.D Degree in 2007 from Anna University Chennai. She has to her credit many International and National journal Publications .Many of her papers were published in IEEE Explore, Elsevier and Springer. Her research interests include Ad Hoc Networks, Wireless Security and Cryptography

called the Central Authority. This leads to an inter level communication mechanism suitable for real-time applications used by the corporate networks .Adding a supervisor distributed in each control and monitoring module allows to manage the multifarious data flow taking part in the inter levels communication.

The major growth of the World Wide Web provides a great deal of potential in supporting cross-platform cooperative work within widely-dispersed working groups**.** The proposed scheme can fair well in such applications**.**

The paper is organized as follows. Section 2 gives a review of previous methods studied for the Hierarchical access control problem. The methodology of the proposed scheme is described in Section 3 and the numerical illustration is discussed in Section 4 with concluding remarks in Section 5

II. RELATED WORK

Several solutions based on cryptographic techniques [1, 2, 12, 14, 3, 10, 17, 20, 16, 19] that address the problem of access control in a hierarchy have been proposed. Most of them employ complex Cryptographic techniques [1, 2, 15, 3, and 17]. Integrating these with existing systems may not be very trivial. Others have undesirable requirements [1, 2, 5, and 14]. Managing key dynamics is a herculean task. When a user joins or leaves a Single Class, it results in an eruption of key changes. The implementation proposed is completely general and can be used to implement different kinds of access control policies of an organization. One of the early solutions to the hierarchical access control problem by Akl and Taylor [1] is based on modular exponentiation. The authors select the exponents in such a manner, that the key of a child node can be derived from the key of its parent. MacKinnon et al. [11] optimized the scheme by preferring less space consuming exponents. However, even for optimized parameters, the public space required remains exponential in the number of nodes. Hwang and Yang [7] presented a scheme that improved the scheme of Akl and Taylor. This scheme is as well based on modular exponentiation, but the innovation is the bottom-up approach. The space complexity remains exponential.

Lin, Hwang and Chang [9] also propose a scheme based on modular exponentiation, bottom-up. This scheme solves the problem of exponential public space requirement. Sandhu [16,17] proposed a key generation scheme for a tree hierarchy. The solution is based on using dissimilar one-way functions to generate the key for each child node in the hierarchy. The one-way function is selected based on the name of the child. When a new child is added, the keys for the ancestors do not have to be recomputed. Hwang [6] proposed a solution for the general poset . The advantage of this approach is its suppleness. Ray et al. [14] present a scheme, for which the key derivation time is constant. This is achieved using modular exponentiation with the same power for all classes, but each class uses a different derivation modulus. The scheme is interesting from a theoretical point of view. However, the efficiency problem is pushed to private space, which becomes exponential. This means, that users themselves must store huge keys.Das et al. [4] proposed a scheme based on polynomial interpolation. A secret key derivation phase requires the user to interpolate a polynomial associated with a node. The space complexity is also quadratic. Zhong [20] and Lin [8] proposed schemes not based on public key cryptography. The public parameters are associated with edges of the graph instead of nodes. Therefore the public space required is quadratic. Symmetric polynomials have been proposed and used for dynamic conferencing schemes [21,22] and for dynamic hierarchies[24]. Symmetric Polynomial based ECC was used for enhancing security [13]. Many authors [6 ,7, 8] have proposed the Tree-Based Group Diffie-Hellman (TGDH) protocol, wherein each user maintains a set of keys arranged in a hierarchical binary tree [18,25,26]. The TGDH can be used for having a secure group communication. In the proposed scheme TGDH is used in the respective classes rather than global class and the different classes are linked with the use of symmetric polynomial .

III. Design Principles

The Hierarchical Access Control Problem was analyzed and the following design principles were identified for implementing the system.

a. Sharing: Dynamic Privilege management and access control technology must operate effectively in a shared manner to provide class dynamics and local key formation.
b. Dynamic restraints**:** The Access Control system should be able to support fast modification of credentials .Constraints such as confidentiality requirements should be satisfied.
c. Policy Evaluation : The software that checks for privileges in each security class must have well defined termination properties . It is not acceptable if the software potentially requires unbounded computation interruption. The validity check of each user should be instantaneous.
d. Credential and Principal Grouping: For ease of management there should be a possibility to levy policy over a group of Subjects.
e. Environment interaction: Whilst the key functionality of the hierarchical access control should be contained within a well defined software body , it is important that dynamic decisions can be made on the basis of conditions outside the access control software. External database interactions with the Central authority should be taken care.
f. Loose Coupling : The hierarchical access control software and the messages or resources being protected must not be very greatly linked. This will facilitate to upgrade the software , if there is a need.
g. Autonomy : The local administration of the various security classes should be taken care separately and the global control should be exercised for the overall functioning. Hence there is a need for multilevel policy autonomy to effectively manage the Hierarchical Access Control.
h. Self Administration: Some self administration procedures may be included in the Software module of each security class to ease the administration and improving the quality.
i. Audit Policy: The Software module must include options for performing auditing to check any malicious intentions.

IV. Proposed Scheme

The proposed Hierarchical Access Control (HAC) algorithm mainly focuses on establishing Inter -class communication where senders are in a different security class and messages they transmit need to be send to all the users who have higher authority and they may belong to different security classes. The mechanism uses a well defined Software Agent called as the Trusted Intermediary Agent to relay the messages to all the users of the superior classes. To establish the link two different encryptions take place Symmetric polynomial key encryption and TGDH key encryptions.

The symmetric polynomial is a polynomial which gives the same value for different permutations of the variables. The symmetric polynomial (1) is used.

$$P\left(x_1,x_2,\ldots,x_m\right)=\sum_{i_1=0}^{t}\sum_{i_2=0}^{t}\cdots\sum_{i_m=0}^{t}a_{i_1,i_2,\ldots,i_m}x_1^{i_1}x_2^{i_2}\cdots x_m^{i_m}\left(\operatorname{mod}P\right) \tag{1}$$

This feature is exploited in deriving the key of the descendant classes as explained in the algorithm. For establishing the Class key CK_x, Tree Based Group Diffie-Hellman Scheme , is used.

Algorithm

Step 1: To establish a confidential communication within every class SC_x, every employee belonging to SC_x should maintain a secret key CK_x to be used as an encryption/decryption key for his class. This key is formed with the help of a Tree Based group Diffie-Hellman key exchange within the class.

Step 2: Each Class SC_x consists of a Class Controller (CC_x) and a Trusted Intermediary Agent (TIA_x) . The Class Controller is a user who initiates the key refreshment whenever there is a employee change. Usually the employee who joins last is designated as the Class Controller. The Trusted Intermediary is a software agent. The expected functionality of the software agent is discussed under design principles in the previous section

Step 3: The Trusted Intermediary possess two keys CK_x and $SP_{x.}$. The SP_x is formed by using a symmetric polynomial function (2) in association with a central authority by using the following formula.

$$SP_x=g_i(s_1',s_2'\ldots,m-m_i-1)=P(\mathrm{s}_\mathrm{i},\mathrm{s}_\mathrm{i1},\mathrm{s}_\mathrm{i2},\mathrm{s}_\mathrm{im},s_1',s_2'\ldots,m-m_i-1) \tag{2}$$

Where s_i are numbers associated with SC_i for i = 1,2,…n and (m-1) additional random numbers s_j ′ for j = 1,2,… m-1 . s_i and s_j belong to Z_p . m is a parameter indicating the number of levels that may be supported by the hierarchy.

Step 4: The Class Dynamics is taken care by the Central Authority.

Step 5: All messages Msg_x are transmitted within a class using their respective CK_x. They are decrypted and seen by the users of the respective class including the TIA_x. The TIA_x now encrypts the message using SP_x and multicasts it.

Step 6: TIA's belonging to Upper Classes SC_u will derive the key SP_x using the symmetric polynomial Approach as given in (3) .

$$S_{j/i}=S_j/(S_iU\left\{SC_i\right\})=\{\mathrm{S}_{(\mathrm{j/i})1},\mathrm{S}_{(\mathrm{j/i})2},\mathrm{S}_{(\mathrm{j/i})3},\cdots\mathrm{S}_{rj}\} \tag{3}$$

Where r_j = | $S_{j/i}$ | and (j/i) r_j is an ordinal number $1\leq$ (j/i $)_d \leq$ n for d = 1,2,…, $r_{j.}$ The set $S_{j/i}$ is a collection of anscestors for class SC_j but excluding SC_i and those classes who are ancestors of both classes SC_i and $SC_{j.}$ Consider a security class

SC_i which is ancestor to security class SC_j and key SP_j can be calculated by SC_i using (4) as,

$$SP_{j(derivedbyi)} = g_i(s_j, s_{(j/i)1}, s_{(j/i)2}, \ldots s_{(j/i)rj}, s'_1, s'_2 \ldots, s'_{m-m_i-2-rj})$$
$$= P(s_i, s_j, s_{i1}, s_{i2}, s_{imi}, s_{(j/i)1}, s_{(j/i)2}, \ldots s_{(j/i)rj}, s'_1, s'_2 \ldots, s'_{m-m_i-2-rj}) \quad (4)$$

Since s_i and s_j' are publicly known , TIA of Class SC_i can compute its key and its descendant key but not its ancestor keys using the polynomial function assigned to Class SC_i. If it tries to calculate the ancestor class key, either the polynomial will give an incorrect key or the function will fail due to disparity in parameters. They then encrypt the message using CK_u and transmit among the users of SC_u.

Step 7. The Users of SC_u will decrypt the message using CK_u.

The proposed scheme satisfies the need for a new simple techniques that is able to provide basic functionalities of satisfying hierarchical access control on the simplest devices and at the same time they can be extended to support advanced functionalities on network which also include high performance infrastructure and other high capability devices . Thus, an adaptive and modular approach is proposed for solving the Hierarchical Access Control
Problem is proposed.

User Dynamics

Member Join Event

The re-keying operation when a new user joins is as follows.

Step 1: The New User gives the Join Request to the Class Controller.
Step 2: The Class Controller changes its contribution and sends the public key to entire employees in the class.
Step 3: Each Employee puts its contribution and computes the New Class Key.
Step 4: The Class Controller sends the key to the Trusted Intermediary agent.
Step 5: The message is encrypted using the Class Key CK_x.

TGDH protocol is used for maintaining the key of the respective classes . The same is done by all the classes. Here a sample illustration for a single class with four users is discussed. Each employee contributes the partial key to compute the class key. This example shows how the shared key is obtained by the employees M_1, M_2, M_3, and M_4. In the Security class initially two employees M_1 and M_2 are available. If the employee M_3 wants to join the security class, it broadcasts join request message to class controller . The class controller receives this message and identifies the insertion point in the tree. The insertion point is the shallowest rightmost node, if an employee joins there, it does not increase the height of the key tree. Otherwise, if the key tree is fully balanced, the new employee joins to the root node. The class controller is the rightmost leaf in the sub tree rooted at the insertion node. When an employee joins in the class, it creates a new intermediate node and promotes the new intermediate node to be the parent of both the insertion node and the new node. After updating the tree, the class controller proceeds to update its share and passes all public keys tree structure to new user. Next, the class controller broadcasts the new tree that contains all public keys. All other employees update their trees accordingly and compute the new class key.

To find the key value of a node:

$K_{<l,v>} = f(K_{<21,2v>} * K_{<21,2v+1>})$
Where
l – level , v – vertices and K – private key
To find the Public key value
$BK_{<l,v>} = f(K)$
$= g^K \text{ mod p.}$
Co-path:
Co-path means the set of siblings in the key path of a node to root.

Initializing the Class:

Initially two Employees M_1 and M_2 are available in the Class. $M_1$and M_2 is going to exchange their keys: Take g = 5 and p = 32713. Employee M_1's private key is 79342, so M_1's public key is 16678. Employee M_2's private key is 85271, so M_2's public key is 27214.

$BK_{<1,0>} = f(K_{<1,0>}) = f(79342)$
$= g^{K_1} \bmod p = 5^{79342} \bmod 32713 = 16678$
$BK_{<1,1>} = f(K_{<1,1>}) = f(85271)$
$= g^{K_1} \bmod p = 5^{85271} \bmod 32713 = 27214$

The Class key is computed (Figure 3) as employee M_1 sends its public key 16678 to employee M_2, M_2 computes their classkey as 12430. Similarly, M_2 sends its public key 27214 to M_1, and then M_1 computes their outer group key as 12430.
Employee node <1,0>
Co-path = { $BK_{<1,1>}$, $K_{<1,0>}$ }
$GK = BK_{<1,1>}^{K_{<1,0>}} \bmod p$
$= 27214^{79342} \bmod 32713 = 12430$
Employee node <1,1>
Co-path = { $BK_{<1,0>}$,$K_{<1,1>}$}
$GK = BK_{<1,0>}^{K_{<1,1>}} \bmod p =$
$16678^{85271} \bmod 32713 = 12430$

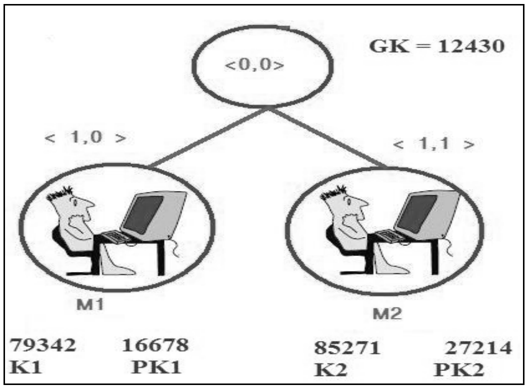


Figure. 1. Employee M_1 and M_2 join the Class

Employee M_3 joins the Class:
When new employee M_3 joins the class, the class controller passes the public key value of tree to employee M_3.

$BK_{<2,1>}$ = f ($K_{<2,1>}$).

Then, M_3 generates the public key from its private key and computes the Class as shown in Figure 4.

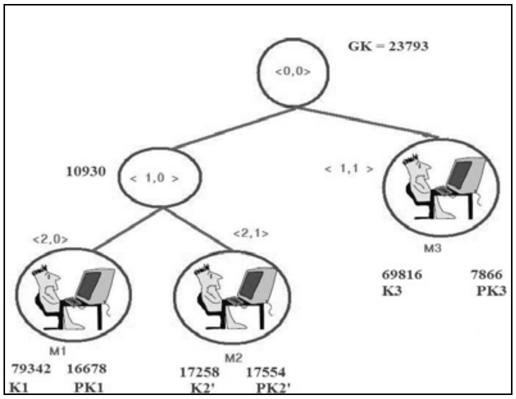


Figure 2 Employee M_3 Joins the Class

Employee Leaving a class:

When a member leaves the class to which it belongs, Class Key of that security class must be changed to preserve the forward secrecy If an employee wants to leave the class, first it should send the leave request to the class controller to generate the new class key. When the leave request message is received by class controller, it updates its key tree by deleting the leaf node corresponding to leave employee. The former sibling of outgoing employee is promoted to parent node. The class controller generate a new private key share, computes all public key pairs on the key-path up to the root and broadcasts the new key tree that contains all public keys. The entire employees in the class compute the new class key. It is assumed to use a constructor function to create the class controller as the first employee which will be existing throughout the HAC.

Employees may leave the session and When an employee M_3 leaves (Figure 5) the Class, then the Class Controller changes its private key and outer group key is recalculated. Now the employee M_4 will move to the level (1,1).

After that, it broadcasts its public key value of tree to all employees in the group. Then, the new class key will be generated by the remaining employees.

Employee Node <1,1>

Co-path = $\{BK_{<1,0>}, K_{<1,1>}\}$

$$= BK_{<1,0>}^{K_{<1,1>}} \bmod 32713 =$$

$$10930^{55181} \bmod 32713 = 13151$$

Employee Node <2,0>

Co-path = $\{BK_{<1,1>}, BK_{<2,1>}, K_{<2,0>}\}$

GK

$$= ((BK<1,1>)^{((BK<2,1>)^{(K<2,0>)}} \bmod 32713) \bmod 32713$$

$= 13151$

Employee Node <2,1>

Co-path = $\{BK_{<1,1>}, BK_{<2,0>}, K_{<2,1>}\}$

$$= ((BK<1,1>^{((BK<2,0>)^{(K<2,1>)}} \bmod 32713)) \bmod 32713$$

$= 13151$

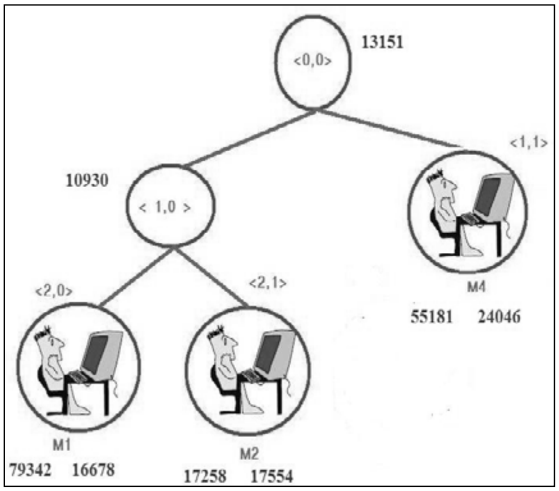


Figure 3. Employee M_3 leaves the Outer group

The Class Key is transmitted to the Trusted Intermediary agent.

Class Dynamics

The handling of Class Dynamics using symmetric polynomial approach is explained below

Adding of a Security class

When a new security class SC_r is added, we need to verify whether m value satisfies the new node restrictions

1) If m < max {m_1, m_2... m_n,..m_r} + 1, a new m value should be used so that m ≥ max {m_1, m_2, ..., m_n,..m_r} + 1. Also, the Central Authority will kindle a new polynomial functions $P\left(x_1, x_2, \ldots, x_m\right)$ accordingly.

2) If m ≥ max{m_1, m_2, ..., m_n,… m_r} + 1, the Central Authority selects a random number s_r for the new security class SC_r so that a new polynomial function g_r can be computed and transmitted to security class SC_r securely. However, if security class SC_r is added as a parent security class of any existing security classes, we need to modify keys of SC_r's descendant security classes to prevent security class SCr from obtaining old keys of its descendant.

Deleting a Security class

When a security class SC_r is detached from the hierarchy, we need to decide whether the security class SC_r is a leaf node or a parent node. Here, a leaf node a node without any descendant:

1)Security class SC_r is a leaf node: The Central Authority can obviously get rid of the public parameter s_r without changing any other keys.

2) Security class SC_r is a parent node: Once security class SC_r is deleted from the hierarchy, we cannot allow it to calculate keys of SC_r's descendant security classes using polynomial function g_r. We need to thwart security class SC_r from accessing its descendants' resources.

Moving of a Security class

A security class SC_r can be moved from one node to another node in the hierarchy. There are four cases:

1) Leaf node to another leaf node: the Central Authority simply re-computes new polynomial function g_r according the new hierarchy and securely transmits g_r to SC_r.

2) Leaf node to parent node: the Central Authority re-computes polynomial functions of security class SC_r and SC_r's new descendant security classes according to the new hierarchy. The Central Authority securely transmits polynomial functions to the affected security classes;

3) Parent node to leaf node: the Central Authority re-computes polynomial functions of previous descendant security classes of SC_r and security class SC_r according to the new hierarchy and then, securely transmits these polynomial functions to the affected security classes

4) Parent node to parent node: the Central Authority re-computes polynomial functions of previous and present descendant security classes of SC_r and security class SC_r according to the new hierarchy and then, securely transmits these polynomial functions to the affected security classes.

Merging of a Security class

Two or more security classes can merge together and become one security class SC_r. Similarly, the Central Authority needs to find previous and present descendant security classes of the merging security classes. The Central Authority randomly takes a new number s_r and then, generates polynomial functions for all corresponding security classes.

Splitting of a Security class

A security class SC_r splits into two security classes SC_{r1} and SC_{r2}. Depending on whether SC_r is a parent node or leaf node, the Central Authority has to decide what previous and present descendant security classes are associated with these security classes (SC_r, SC_{r1} and SC_{r2}). The Central Authority then selects two new numbers s_{j1} and s_{j2} and generates polynomial functions for these affected security classes.

Adding a Link in the Hierarchy

If two security classes SC_r and SC_k are linked together, we establish a new direct parent-child relationship between two security classes, say security class SC_r is the parent of security class SC_k. There are two different cases: 1) security class SC_r was an ancestor of security class SC_k through other security classes. The Central Authority does not need to perform anything; and 2) security class SC_r is the only parent for security class SC_k in the new hierarchy. The Central Authority selects a new number S_k, and generates new polynomial functions for security class SC_k and its descendants security classes. The Central Authority securely transmits new polynomial functions to these affected security classes.

Deleting a Link

If two linked security classes SC_r and SC_k are disconnected, the direct parent-child relationship is removed between two security classes. Security class SC_r will not be the parent of security class SC_k in the fresh hierarchy. Again, there are two different cases: 1) security class SC_r is still an ancestor of security class SC_k through other security classes in the new hierarchy. The Central Authority does not need to perform anything; and 2) security class SC_r is not an ancestor for security class SC_k in the new hierarchy. The Central Authority selects a new number s_k, and generates new polynomial functions for security class SC_k and its descendant security classes. The Central Authority securely transmits new polynomial functions to these affected security classes

V. References.

1. S. G. Akl and P. D. Taylor. Cryptographic Solution to a Multilevel Security Problem. In D. Chaum, R. L. Rivest, and A. T. Sherman, editors, *Advances in Cryptology: Proceedings of CRYPTO '82*, pages 237–249. Plenum Press, NY, August 1982.

2. S. G. Akl and P. D. Taylor. Cryptographic Solution to a Problem of Access Control in a Hierarchy. *ACM Transactions on Computer Systems*, 1(3):239–248, 1983.

3. G. C. Chick and S. E. Tavares. Flexible Access Control with Master Keys . In G. Brassard, editor, *Advances in Cryptology: Proceedings of Crypto '89*, volume 435 of *Lecture Notes in Computer Science*, pages 316–322. Springer-Verlag, 1990.

4. Manik Lal Das , Ashutosh Saxena , Ved P. Gulati, and Deepak B. Phatak. Hierarchical key management scheme using polynomial interpolation. SIGOPS Oper. Syst. Rev., 39(1):40 - 47, 2005.

5. W. Diffie and M. E. Hellman. New directions in cryptography. IEEE Transactions on information theory, IT-22(6):644{654, Nov 1976.

6. Min-Shiang Hwang. A new dynamic key generation scheme for access control in a hierarchy. Nordic J. of Computing, 6(4):363{371, 1999.

7. Min-Shiang Hwang and Wei-Pang Yang. Controlling access in large partially ordered hierarchies using cryptographic keys. J. Syst. Softw., 67(2):99 - 107,2003.

8. Chu-Hsing Lin. Hierarchical key assignment without public-key cryptography. Computers and Security, 20(7):612{619, 2001.

9. Iuon-Chang Lin, Min-Shiang Hwang, and Chin-Chen Chang. A new key assignment scheme for enforcing complicated access control policies in hierarchy. Future Generation . Computing . Systems., 19(4):457{462, 2003.

10. H. Y. Lin and L. Harn A Cryptographic Key Generation Scheme for Multi-level Data Security. *Computer & Security*, 9(6):539–546, 1990.

11. S. J. MacKinnon, P. D. Taylor, H. Meijer, and S. G. Akl. An Optimal Algorithm for Assigning Cryptographic Keys to

Access Control in a Hierarchy. *IEEE Transactions on Computers*, C-34(9):797–802, 1985

12. Rolf H. Mohring. Computationally tractable security classes of ordered sets. Algorithms and Order (I. Rival, ed.), pages 105 {183, 1989}

13. Nafeesa Begum.J, Kumar.K,Sumathy.V, A Novel Approach towards Multilevel Access Control for Secure Group Communication Using Symmetric Polynomial Based Elliptic Curve Cryptography, International Conference on Computational Intelligence and Communication Networks, p. 454-59, November 2010.ISBN: 978-0-7695-4254-6

14. Indrakshi Ray, Indrajit Ray, and Natu Narasimhamurthi. A cryptographic solution to implement access control in a hierarchy and more. In SACMAT '02: Proceedings of the seventh ACM symposium on Access control models and technologies, pages 65{73. ACM Press, 2002.

15. R. L. Rivest, A. Shamir, and L. Adleman. A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. *Communications of the ACM*, 21(2):637–647, 1978.

16. R. S. Sandhu. Cryptographic Implementation of a Tree Hierarchy for Access Control. *Information Processing Letters*, 27(2):95–98, 1988

17. R. S. Sandhu. On some cryptographic solutions for access control in a tree hierarchy. In ACM '87: Proceedings of the 1987 Fall Joint Computer Conference on Exploring technology: today and tomorrow, pages 405{410. IEEE Computer Society Press, 1987.

18. W.H.D., Howarth, M., Sun, Z. and Cruickshank, H. “Dynamic balanced key tree management for secure multicast communications”, IEEE Transactions on Computers, 56(5), pp.577-589, 1997

19. Y. Zheng, T. Hardjono, and J. Seberry. New Solutions to the Problem of Access Control in a Hierarchy. Technical Report Preprint 93-2, Department of Computer Science, University of Wollongong, February 1993.

20.Sheng Zhong. A practical key management scheme for access control in a user hierarchy. Computers & Security, 21(8):750{759, 2002.

21. Zou.X , Ramamurthy.B &. Magliveras.S, Secure Group Communications over Data Networks, Springer, New York, NY, USA, ISBN: 0-387-22970-1, October,2001.

22. C. Blundo, A. De Santis, A. Herzberg, S. Kutten, U. Vaccaro, M. Yung, Perfectly Secure Key Distribution for Dynamic Conferences. In *Advances in Cryptology-CRYPTO’92*, LNCS, 740 (1993), pp.471–486.

23. R. Aparna, and B. B. Amberker, Analysis of Key Management Schemesfor Secure Group Communication and Their Classification, Journal of Computing and Information Technology - CIT 17, 2009, 2, 203–214.

24. X.Zou and L.Bai , A New Class of Key Management Scheme for Access Control in Dynamic Hierarchies, “ International Journal of Key management Scheme for Access Control in Dynamic Hierarchies, International Journal of Computers and Applications, Vol. 30, No.4, 2.

25.Dutta, R., Barua, R. and Sarkar, P. “Provably secure authenticated tree based group key agreement”, in Proc. ICICS 2004, Lecture Notes in Computer Science, Vol. 3269, pp.92–104, 2004.

26.Kim, Y., Perrig, A. and Tsudik, G. “Tree-based group key agreement”, ACM Transactions on Information Systems security, 7(1), pp. 60-96, 2004.

Dynamic Adaptive Streaming over HTTP (DASH) using feedback linearization: a comparison with a leading Italian TV operator

Vito Caldaralo, Luca De Cicco, Saverio Mascolo, and Vittorio Palmisano

Abstract— **Dynamic Adaptive Streaming Over HTTP (DASH) is a video streaming standardization effort that aims at building video content delivery systems that dynamically adapt video bitrates to match the time-varying bandwidth of a conventional *HTTP* connection over the Internet Protocol. In particular, the video content is segmented into a sequence of chunks, each one containing a short interval of playback time, for instance 6 seconds. Chunks are encoded at different bit rates and a DASH client, at the receiver side, automatically selects the next chunk to download based on current Internet available bandwidth. The control goal is to provide the best possible video quality while avoiding playback video interruptions, i.e. rebuffering events, in the presence of an unpredictable bandwidth. In this paper we design a new DASH controller using feedback linearization. A real implementation of the controller is tested and compared with the automatic video switching control of a leading Italian TV operator.**

I. Introduction and related work

Recent developments in Future Internet platform development are opening several possibilities for new multimedia applications [4]. The video part of Internet traffic is booming due to the ever increasing availability of video content from site such as YouTube (video sharing), Netflix (movie on demand), Livestream (live streaming) and the pervasive diffusion of Tablets, Smart-Phones and SmartTVs which access the Internet through broadband wired and wireless links. The Cisco Visual Networking Index predicts that video will be 69 percent of all consumer Internet traffic in 2017 excluding peer-to-peer (P2P) video file sharing, whereas the sum of all forms of video (TV, video on demand and P2P) will be in the range of 80 to 90 percent of global consumer traffic by 2017 [5].

The key technological choice that ignited the start of video distribution over the Internet at large scale was the use of the HTTP protocol over the TCP. This choice was done by YouTube in 2005. The first approach to video streaming was the *progressive download streaming* where a video was encoded at a given quality and sent to the user as any other file using a HTTP connection. This approach has the following main problem: the video is encoded at a given bitrate, which is not elastic, but it is transported through the TCP that, on the contrary, is designed for elastic traffic over best-effort Internet. As a consequence, a persistent mismatch between the video bitrate and the best-effort Internet available bandwidth may result in an empty video playout buffer with video interruptions.

For this reason it is necessary to make the video content adaptive at least to some extent. The leading proposed approach consists of encoding a video at different bitrates and resolutions, the *video levels*, and video levels are divided into *segments* or chunks of fixed durations. A stream-switching *controller* at the client selects the next chunk to be downloaded at the best possible quality given the available bandwidth with the constraint of avoiding video interruptions. Standard HTTP servers can be used for video distribution [13] and scalability can be easily obtained using CDNs. Typically, adaptive players work as follows: at the beginning of the connection the player requests the video segments, of fixed duration τ, through consecutive HTTP GET requests in order to build the buffer; then, when a certain amount of video is stored in the playout buffer, the *buffering-phase* is completed and the player enters in the *steady-state* phase; while in this state, the player strives to maintain the playout buffer level constant by issuing the HTTP requests each τ seconds. Thus, the player generates a on-off traffic pattern during the steady-state: the video segments are downloaded during the ON phase and then, during the OFF phase, the player remains idle until the next download is started [10], [1], [3].

It has been shown that the client-side algorithms proposed so far generate an on-off traffic pattern at steady-state that can lead to unfairness when many video flows share the same bottleneck [1], [2], [10]. Moreover, [8] shows that three popular video on demand streaming services in the US, Netflix, Vudu and Hulu, are not able to obtain their fair share of bandwidth in the presence of coexisting TCP greedy flows. This phenomenon was called "*downward spiral effect*" and was explained as an effect of the on-off traffic pattern generated at the sender side.

To overcome the before mentioned issues, several adaptive streaming algorithms have been proposed. FESTIVE has been proposed to provide fairness in a multi-client scenario [9]. PANDA [11] computes the chunk inter-request time to provide fairness and cut video bitrate oscillations. [2] proposes to introduce a traffic shaper at the server to eliminate the OFF phases when the player is in steady-state.

In [6] it has been shown that the automatic stream-switching system of a major CDN operator employs a different approach *wrt* the classic client-side architecture. In particular, it employs a hybrid sender-side/client-side architecture with two controllers running at the client: one for

L. De Cicco, V. Palmisano, and S. Mascolo are with the Dipartimento di Ingegneria Elettrica e dell'Informazione at Politecnico di Bari, Via Orabona 4, 70125, Bari, Italy Emails: l.decicco@poliba.it, vpalmisano@gmail.com, mascolo@poliba.it.

V. Caldaralo is with CRAT Bari, Via Orabona 4, 70125, Bari, Italy, Email: vito.caldaralo@gmail.com

selecting the video level, the other for throttling the sending rate at the server in order to control the playout buffer length at the client. Moreover, the system does not issue a lot of HTTP GET requests to download the segments, but sends HTTP POST requests to the server to select the video level to be streamed.

In this paper we propose a novel client side controller, named ELASTIC (*fEedback Linearization AdaptIve STreamIng Controller*), that has been designed using feedbeack linearization. The proposed control algorithm is able to avoid the on-off traffic pattern at the sender and to get its fair share of the bottleneck bandwidth when coexisting with TCP greedy flows. Finally we develop an experimental comparison of ELASTIC with a leading Italian TV operator. In particular, to compare the considered algorithms, we have set up a controlled testbed that allows bandwidth and delays be set.

The paper is organized as follows: in Section II ELASTIC is presented; Section III describes the employed testbed; Section IV presents the results of the experimental evaluation and Section V concludes the paper.

II. ELASTIC

In this section we propose *ELASTIC*, a client-side adaptive streaming algorithm designed using feedback control theory. In Section II-A we present the design requirements; Section II-B describes the control system model; in Section II-C the control algorithm is presented, and Section II-D provides the controller implementation details.

A. Design requirements

The main goal of a stream-switching controller is to dynamically select the *video level* $l(t) \in \mathscr{L} = \{l_0, \dots, l_{N-1}\}$ for each video segment to achieve the maximum Quality of Experience (QoE) while avoiding video interruptions that happen when the receiver playout buffer gets empty. Re-buffering events, occurring when the player buffer gets empty, have been identified to be one of the major causes impairing user engagement [7]. Moreover, it has been shown that frequent quality switches may be annoying to the user [12], thus limiting video level switches is considered a design requirement by several proposed algorithms [9], [11].

Summarizing, we consider the following design goals for ELASTIC: 1) minimize the re-buffering ratio; 2) maximize the obtained video level; 3) provide fair sharing of the bottleneck when coexisting with other video or long-lived TCP flows.

B. The control system model

Figure 1 shows the block diagram of a DASH streaming system with the controller at the client-side. The HTTP server sends the video to the client through an Internet connection with an end-to-end bandwidth $b(t)$ and a round-trip-time (RTT) equal to T. The client receives the video segments at a rate $r(t) < b(t)$, and temporarily stores them in a *playout buffer* that feeds the video player. The *controller* dynamically decides, for each video segment, the video level $l(t)$ to be downloaded sending a `HTTP GET` request to the HTTP server. The *measurement* module feeds the controller with measurements such as the estimated bandwidth $\hat{b}(t)$ and the playout buffer level $q(t)$.

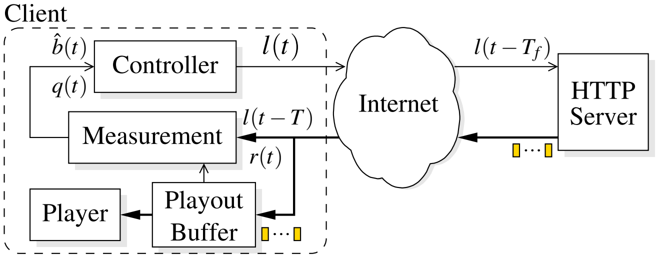


Fig. 1. Client-side adaptive video streaming

C. The adaptive streaming controller

The typical approach to implement a stream-switching system is to design two controllers [3], [6], [8], [11]: one throttles the video level $l(t)$ to match the measured available bandwidth $b(t)$, the other regulates the playout buffer length $q(t)$ by controlling the idle period between two segment downloads.

Differently from the currently used approaches, ELASTIC uses a unique controller that selects the video level $l(t)$ to drive $q(t)$ to a set-point q_T. This eliminates the idle periods between segment downloads. Indeed, by reaching the controller goal, i.e. $q(t) \to q_T$, the video level $l(t)$ also matches the available bandwidth $b(t)$, i.e. the maximum possible video level is obtained.

The playout buffer length $q(t)$, i.e. the seconds of video stored in the playout buffer, can be modelled as an integrator:

$$\dot{q}(t) = f(t) - d(t),$$

where $f(t)$ is the *filling rate* and $d(t)$ is the *draining rate*.

Let us focus on the filling rate, which is equal by definition to dt_v/dt, where dt_v is the amount of video duration received by the client in a time dt. The *video encoding bitrate* is defined as $l(t) = dD/dt_v$, where dD is the amount of bytes required to store a portion of video of duration dt_v. It is important to notice that $l(t)$ is always strictly greater than zero by definition. The *received rate* $r(t)$ is defined as $r(t) = dD/dt$, i.e. the amount dD of bytes received in a time interval dt. Thus, since $f(t) = dt_v/dt = (dt_v/dD) \cdot (dD/dt)$, it turns out that:

$$f(t) = \frac{r(t)}{l(t)}. \tag{1}$$

We now derive the model of the draining rate $d(t)$. The playout buffer is drained by the player: when the video is playing, dt_v seconds of video are played in $dt = dt_v$ seconds, i.e. $d(t) = 1$; on the other hand, when the player is paused the draining rate is zero. Thus, $d(t)$ can be modelled as follows:

$$d(t) = \begin{cases} 1 & \text{playing} \\ 0 & \text{paused} \end{cases} \tag{2}$$

```
1: On segment download:
2: $\Delta T \leftarrow getDownloadTime()$
3: $S \leftarrow getSegmentSize()$
4: $d \leftarrow isPlaying()$
5: $q \leftarrow getQueueLength()$
6: $r \leftarrow h(S/\Delta T)$
7: $q_I \leftarrow q_I + \Delta T \cdot (q - q_T)$
8: return $Quantize(r/(d - k_p q - k_i q_I))$
```

Fig. 2. ELASTIC controller pseudo-code

Finally, by combining (1) and (2) we obtain the playout buffer length model:

$$\dot{q}(t) = \frac{r(t)}{l(t)} - d(t). \tag{3}$$

The video level $l(t)$ is the control variable, $q(t)$ is the output of the controlled system, whereas $r(t)$ can be modelled as a disturbance. It is important to notice that $l(t)$ can only assume values in the discrete set $\mathscr{L}$, i.e. the output of the controller is quantized.

In the following we employ the feedback linearization technique to compute a control law that linearizes (3) and that steers $q(t)$ to the set-point q_T. To this end, we impose the following linear closed-loop dynamics for the queue:

$$\dot{q}(t) = -k_p q(t) - k_i q_I(t) \tag{4}$$

$$\dot{q}_I(t) = q(t) - q_T \tag{5}$$

where q_I is an additional state that holds the integral error, $k_p \in \mathbb{R}_+$ and $k_i \in \mathbb{R}_+$ are the two parameters of the controller.

Now, by equating the right-hand sides of (3) and (4), it turns out:

$$l(t) = \frac{r(t)}{d(t) - k_p q(t) - k_i q_I(t)} \tag{6}$$

that is the control law employed by the stream-switching controller.

D. Implementation

Figure 2 shows the pseudo-code of the controller. When a segment is downloaded, the following quantities are measured: 1) the time spent to download the segment ΔT (line 2); the last downloaded segment size S in bytes (line 3); state of the player d (line 4); the playout buffer length (line 5); the received rate r is estimated by passing the last segment download rate $S/\Delta T$ through a harmonic filter $h(\cdot)$ over the last 5 samples of r (line 6). Then, the integral error q_I is updated (line 7) and the control law is computed using (6) (line 8).

III. Testbed

In this Section we describe the testbed, the experimental scenarios, and the metrics employed to evaluate and compare ELASTIC with the automatic video switching control of a leading Italian TV operator.

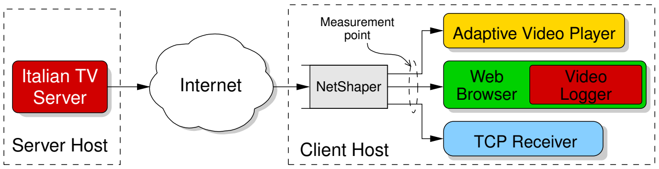


Fig. 3. Testbed setup.

A. The testbed

Figure 3 shows the testbed that we have employed to carry out the experimental evaluation: the receiving host, or client, is a Debian Linux machine connected to the Internet via our 100 Mbps campus wired connection. The `Adaptive Video Player (AVP)`, which we have developed in Python, runs on it. AVP is implemented using the GStreamer[1] libraries and supports playback of MP4-encoded videos under the control of ELASTIC as adaptive streaming algorithm. Moreover, one or more Google Chrome Web Browser instances runs on the client with an embedded *Video logger extension* developed by us to measure the adaptive streaming metrics of the automatic video switching control considered in the comparison. The *Video logger extension* parses the video playlists, captures the video chunks download requests and measures both the playout buffer length and the portion of buffered video. We have always used the web browser in incognito mode to prevent from interference with caching techniques by removing data stored in the browser after previous views.

In order to set the bottleneck bandwidth capacity and propagation delays we have developed `NetShaper`. This tool employs the `nfqueue` library provided by `Netfilter`[2] to capture and redirect to a user space drop-tail queue the packets arriving at the client. The traffic shaping policies are performed on this queue.

Before running each experiment, we have carefully checked that the end-to-end available bandwidth between the TV operator's server and the client was well above the bottleneck capacity set by the traffic shaper. It is worth noting that all the measurements we report in the paper have been performed after the traffic shaper.

In each experiment we have used the same video sequence, encoded at two different bitrates as shown in Table I, with resolutions 700x394 and 1024x574. The duration of each chunk is 6s.

For each video, both the player and the *Video logger extension* are able to log: 1) the playout buffer length $q(t)$ measured in seconds, 2) the video level $l(t)$, 3) the cumulative downloaded bytes $D(t)$, 4) the cumulative re-buffering time $T_{rb}(t)$, 5) the number of re-buffering events $n_{rb}(t)$, 6) the number of level switches $n_s(t)$.

B. Scenarios and metrics

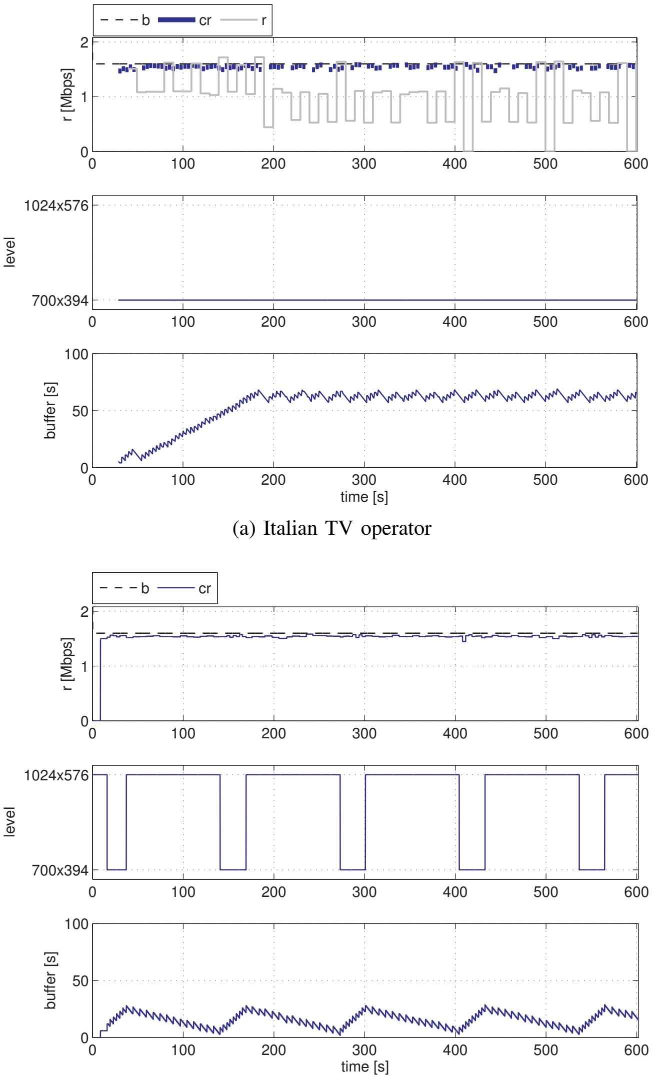
We have considered three scenarios: (S1) one video over a bottleneck link whose available bandwidth is set to

[1] http://gstreamer.freedesktop.org/
[2] http://www.netfilter.org/

TABLE I

DISCRETE SET OF VIDEO LEVELS $\mathscr{L}$.

Video level	l_0	l_1
bitrate (kbps)	844	1845
Resolution	700x394	1024x576



(a) Italian TV operator

(b) ELASTIC

Fig. 4. One video over a 1.6Mbps bottleneck link.

$b = 1.6$Mbps; (S2) one video over a square-wave varying bottleneck link with a period of 200 s, a minimum value $A_m = 1$ Mbps and a maximum value $A_M = 4$ *Mbps*; this scenario is aimed at showing the dynamic behaviour of the considered algorithms; (S3) One video sharing a bottleneck link whose available bandwidth is set to $b = 4$ Mbps with one long-lived TCP flow.

For each scenario we will show the dynamics of the following variables: 1) the chunk download bitrate $cr(t)$; 2) the received video bitrate $r(t) = D(t)/\Delta T$ with $\Delta T = 10$s; 3) the received video level $l(t)$; 4) the playout buffer length $q(t)$ measured in seconds. .

IV. RESULTS

In this Section we provide the details of the results obtained for both the considered algorithms.

A. One video over a 1.6 Mbps link

In this scenario we investigate the dynamic behaviour of an ELASTIC client and an automatic video switching control used by a leading Italian TV operator client to stream a video over a 1.6Mbps bottleneck link. This capacity is lower than the bitrate of the maximum level l_1 and higher than the bitrate of the minimum level l_0. The experiment duration is 600s.

Figure 4 shows the dynamics for each considered metric. The video connections start at $t = 0$ s. The figure shows that ELASTIC obtains a bitrate very close to the maximum possible, with a video level oscillating between l_0 and l_1 since the available bitrate is in the range $[l_0, l_1]$. It is worth noting that the average video level obtained in this experiment is equal to the available bandwidth. On the other hand, in the case of the Italian TV operator the video level is always l_0, indicating that the maxmimum quality is not achieved.

B. The case of a square-wave varying bottleneck capacity

In this scenario we investigate how the quality adaptation algorithm reacts in response to abrupt drops/increases of the bottleneck capacity. Towards this end, we let the bottleneck capacity to vary as a square-wave with a period of 200 s, with a minimum value $A_m = 1$ Mbps and a maximum value $A_M = 4$Mbps. The aim of this experiment is to assess if the considered players are able to quickly change the video level when an abrupt drop/increase of the bottleneck capacity occurs in order to guarantee continuous reproduction of the video content at the maximum quality.

In the first case we have considered an initial bottleneck capacity set to A_M. As shown in Figure 5a, the first result is that the Italian TV player starts with the maximum level due to the fact that the available bandwidth is well above the bitrate of the highest quality level l_1. It makes consecutive HTTP GET requests and fill the playout buffer up to 96 s until $t = 200$ s. At this moment, the available bandwidth goes down, the algorithm reacts immediately with a switch down at level l_0, but it produces an off period while the buffer decrease at 60 *s*, which corresponds to the upper limit for the buffer when the level is l_0. Finally, when there is a new bandwidth increase at $t = 400$ s the controller reacts immediately with another switch up at the highest quality l_1 and produces consecutive chunk requests, as shown by the requests estimated bandwidth *cr*, until the playout buffer reach 146 *s* of video and the test is stopped.

In the figure 5b we can see the dynamic behaviour of the ELASTIC algorithm. The player enters in the steady-state phase when the playout buffer is up to 60 *s*. When the available bandwidth goes down at $t = 200$ s the player does not react immediatly with a switch down although the estimated bandwidth is close to A_m and lower than the bitrate of l_1. This is due to the fact that the playout buffer is sufficient to avoid re-buffering events while the download at l_1 is completed. In the interval between $t = 200$ s and

$t = 400$ s the available bandwidth is higher than the bitrate of l_0 and lower than the bitrate of l_1, so the control action of the ELASTIC algorithm provides some switches to mantain the playout buffer close to the set-point. Finally, when the available bandwidth return to $A_M = 4$Mbps the ELASTIC algorithm mantains the maximum level end reach the steady state.

Now we consider an initial bottleneck capacity set to A_m as shown in Figure 6. When the bottleneck capacity goes up to A_M at $t = 200$ s, the Italian TV player starts with consecutive HTTP GET requests to fill the playout buffer. When the available bandwidth goes down at $t = 400$ s, as in the previous case, the player produces an off period while the buffer decrease at 60 s after which it provides a switch down at level l_0.

In the same scenario, the ELASTIC algorithm provides a very fast tracking of the available bandwidth when occurs the switch up of the bottleneck capacity. When the available bandwidth goes down at $t = 400$ s, as in the previous case, the ELASTIC player mantains the highest video quality until $t = 500$ s despite to the available bandwidth because the playout buffer of 60 s is enough to avoid re-buffering events.

C. One video sharing a 4 Mbps link with one TCP flow

In this scenario we investigate the dynamic behaviour of the considered algorithms when one video and one TCP flow share a 4Mbps bottleneck link. The experiment duration is 600s. In order to obtain a concurrent TCP flow from the same server of the video, we downloaded the maximum level of the same video. The fair share in this experiments is 2Mbps, that corresponds to a bitrate slightly greater than the level l_1 (1.8Mbps). We expect that the level of the video is l_1 for the entire experiment duration.

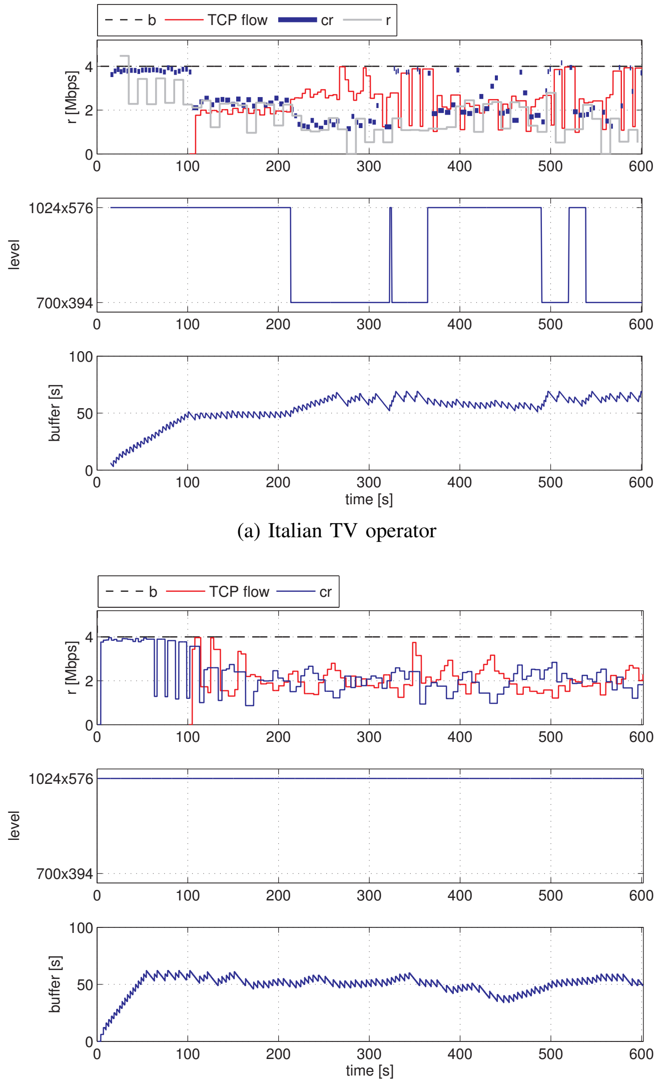
Figure 7 shows the dynamics for both algorithms in the case of a video flow sharing the bottleneck with a long-lived TCP flow starting at $t = 100$s.

The figure clearly shows that the Italian TV player is not able to obtain the fair share when coexisting with a TCP flow and exhibits the "downward spiral effect" shown in [8]. In particular, the steady-state video levels obtained oscillates between l_0 and l_1. On the other hand, ELASTIC obtains a bitrate very close to the fair share, with a video level always at l_1. The ELASTIC flow is able to get the fair share when coexisting with a long-lived TCP flow because it does not produce an on-off traffic pattern thus behaving as a TCP flow.

V. Conclusions

In this paper we have proposed ELASTIC, a novel controller for adaptive video streaming obtained by using feedback linearization. Differently from current existing proposals, ELASTIC uses a unique controller that selects the video level $l(t)$ that drives the playout buffer length to a set-point. ELASTIC eliminates the on-off traffic pattern which causes underutilization and unfairness when video flows coexist with long-lived TCP flows [8].

We have experimentally compared the performance of ELASTIC with the automatic video switching control of a leading Italian TV operator in a controlled testbed. Results have shown that Italian TV player is responsive in changing the video level to match the available bandwidth, but it is not able to grab the fair share when in the presence of coexisting TCP traffic. On the other hand, ELASTIC is always able to get the fair share and to match the available bandwidth on average in all the considered scenarios.



(a) Italian TV operator

(b) ELASTIC

Fig. 7. One video flow sharing a 4Mbps bottleneck link with a TCP flow

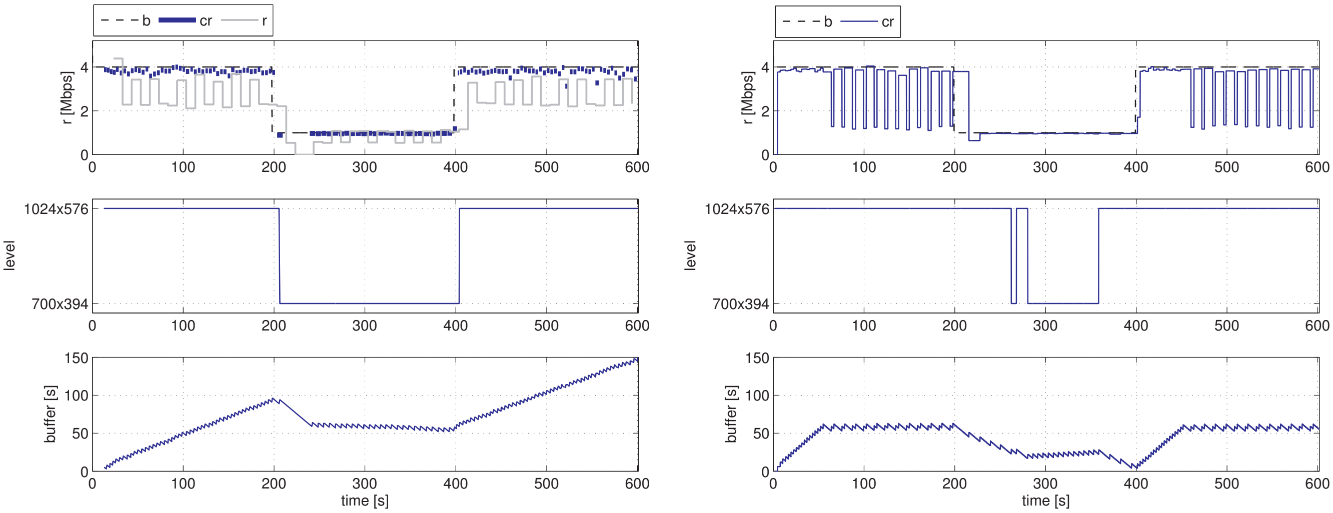
VI. Acknowledgement

This work has been partially supported by the Italian Ministry of Education, Universities and Research (MIUR) through the PLATINO project (PON01 01007).

References

[1] S. Akhshabi, L. Anantakrishnan, C. Dovrolis, and A. C. Begen. What happens when http adaptive streaming players compete for bandwidth? In *Proc. of ACM NOSSDAV '12*, 2012.

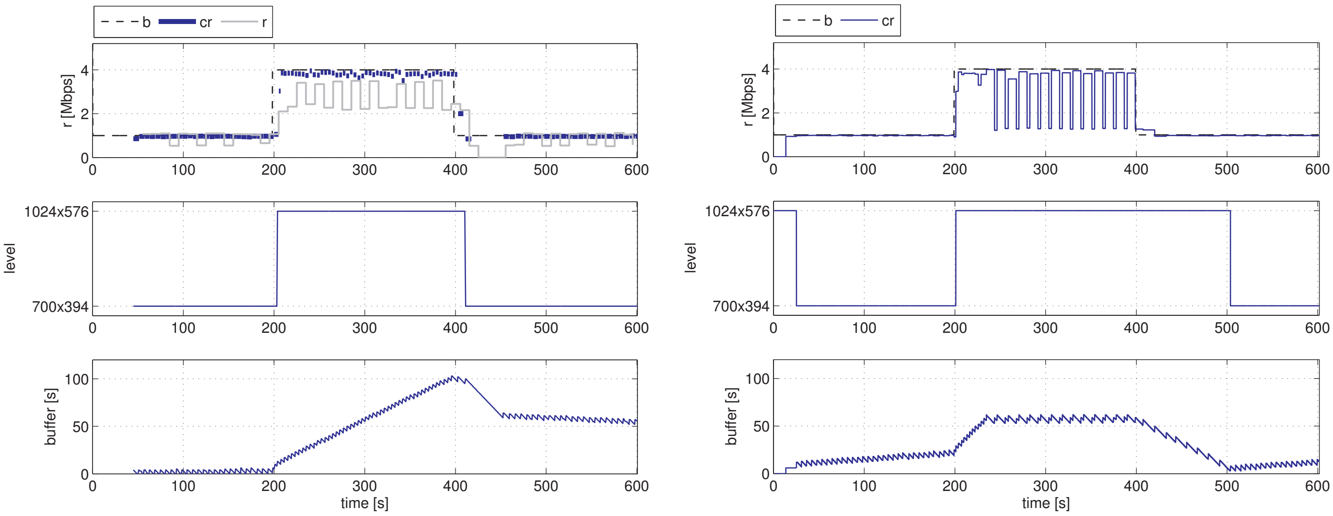
[2] S. Akhshabi, L. Anantakrishnan, C. Dovrolis, and A. C. Begen. Server-based traffic shaping for stabilizing oscillating adaptive streaming players. In *Proc. of ACM NOSSDAV '13*, 2013.



(a) Italian TV operator

(b) ELASTIC

Fig. 5. One video flow sharing a square-wave varying bottleneck link starting with $A_M = 4\,\text{Mbps}$



(a) Italian TV operator

(b) ELASTIC

Fig. 6. One video flow sharing a square-wave varying bottleneck link starting with $A_m = 1\,\text{Mbps}$

[3] S. Akhshabi, A. Begen, and C. Dovrolis. An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP. *Proc. of ACM MMSys 2011*, pages 157–168, 2011.

[4] M. Castrucci, F. D. Priscoli, A. Pietrabissa, and V. Suraci. A cognitive future internet architecture. In *The future internet*, volume 6656, pages 91–102. Springer, 2011.

[5] Cisco. Cisco visual networking index, global mobile data traffic forecast update, 2010-2015. *White Paper*, Feb 2011.

[6] L. De Cicco and S. Mascolo. An adaptive video streaming control system: Modeling, validation, and performance evaluation. *IEEE/ACM Transactions on Networking*, in press.

[7] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang. Understanding the impact of video quality on user engagement. In *Proc. of ACM SIGCOMM 2011*, pages 362–373, 2011.

[8] T. Huang, N. Handigol, B. Heller, N. McKeown, and R. Johari. Confused, timid, and unstable: picking a video streaming rate is hard. In *Proc. of ACM Internet Measurement Conference*, pages 225–238, 2012.

[9] J. Jiang, V. Sekar, and H. Zhang. Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive. In *Proc. of CoNEXT '12*, pages 97–108, 2012.

[10] T. Kupka, P. Halvorsen, and C. Griwodz. Performance of On-Off Traffic Stemming From Live Adaptive Segmented HTTP Video Streaming. In *Proc. of IEEE Conference on Local Computer Networks*, pages 405–413, Oct. 2012.

[11] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Begen, and D. Oran. Probe and adapt: Rate adaptation for http video streaming at scale. *arXiv preprint arXiv:1305.0510*, 2013.

[12] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen. Flicker effects in adaptive video streaming to handheld devices. In *Proc. of 19th ACM international conference on Multimedia*, pages 463–472. ACM, 2011.

[13] I. Sodagar. The mpeg-dash standard for multimedia streaming over the internet. *IEEE MultiMedia*, 18(4):62–67, 2011.

An improved on-the-fly web map generalization process

Brahim Lejdel[1], Okba Kazar[2]

[1] LINFI, University of Biskra , University of el-Oued, Algeria,
[2] LINFI , University of Biskra, Algeria.
lejdel82@yahoo.fr, *kazarokba@yahoo.fr.*

Abstract- Today, great numbers of users want access to spatial data on the web specific to their needs. This may be possible by applying the suitable transformations, in real-time. This process called on-the-fly generalization maps. Many approaches were proposed for improving this process, but those not suffice for guiding a powerful and efficient process. Also, the transformation of topological relationships did not studied in these approaches. In this paper, we will describe mainly the transformation of the topological relationships during the on-the-fly web map generalization. We use mainly, two types of object; Ribbon and regions.

Keywords- On-the-fly maps generalization, topological relationships, multiple representations, Optimization, Genetic Agent.

I. INTRODUCTION

To provide on-the-fly web mapping to the user, the process of on-the-fly map generalization must rely on fast, effective, and powerful methods. A principal challenge of such on-the-fly maps generalization is to offer the user a spatial data in real-time and in height quality. In order to elegantly the map, we have to describe a framework which includes the transformation of spatial relationship between two objects in connection with scale. In this context, we present an efficient and simple process called generalization-reduction-disappearance which was the key to govern the generalization of topological relations [20].

In this paper, we use mainly metric conditions as distance between objects, area and width of objects, to formulate a mathematical framework which composed of a set of assertions for treating the variety of topological relation according the scale.

This paper organized in six sections. Section one presents the definition of on–the-fly web map generalization and of topological relationships. In the section two, the related works were presented. Then, the framework of on the-fly generalization map was described. Section four detailed the transformation of topological relationships. In the section 5, we present the implementation of topological relationships module. Finally, the section 6 present the conclusion and the future work.

II. DEFINITIONS

A. On-The-Fly Web Map Generalization

The on-the-fly web map generalization is defined as; the creation in real-time and according to the user's request, of a cartographic product appropriate to its scale and its purpose, from a largest-scale database. The main characteristics of on-the-fly web mapping are:

- Required maps must be generated in real-time [15].
- Generation of a temporary and reduced scale dataset for visualization purposes from the database [11] in order to use the computer's memory efficiently [4].
- A real-time map generation process has to take into account users' preferences and contexts.
- A real-time map generation process must adapt maps' contents to display space and resolution of display media as well as to the contextual use of these maps [15].
- The scale and theme of the map are not predefined [4].
- There is no way to verify the quality of the final map that will be sent to the user [15].

The main problems linked to on-the-fly map generalization are the time of delivering the cartographic data and its quality. The generalization process time is a crucial factor to provide a user cartographic data. The waiting time must be compatible with Newell's cognitive band, which is less than 10 seconds [13]. Then, to guarantee the best quality of the map, we must study the transformation of topological relationships when downscaling. Also, in order to produce maps suited to a user's requests, on-the-fly map generalization must be flexible enough to take into account the level of detail, the kind of the map [1]…etc.

B. Ribbon

We claim that ribbons may elegantly model rivers and roads (so-called linear objects): a ribbon can be loosely defined as a line or polyline with a width. Mathematically speaking, a ribbon is defined as longish rectangle [20]. The ribbon has a skeleton which is its axis. See Figure 4 for an example.

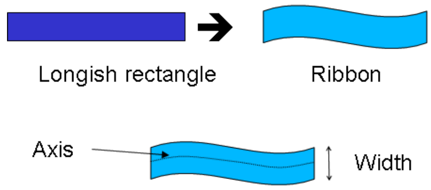


Figure 1. Definition of Ribbon

Let us note *Width(R)* and *Skeleton(R)* respectively the width and the Skelton of a ribbon. Remember that the ribbon can contain holes which can be useful for modeling islands in rivers.

In the sequel of this paper, to simplify the presentation, a ribbon will be represented by a longish rectangle. For instance a motorway (Figure 5) can be described by several ribbons corresponding to several driving lanes, emergency lanes and one median.

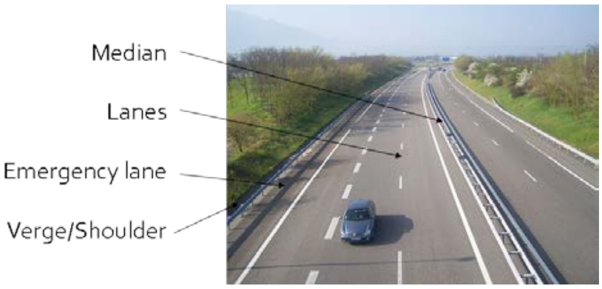


Figure 2. Ribbon model applied to a motorway.

C. **Region**

This feature may represent the building, greene zone and all areal objects. We can deffined a region as Loose polygonal type. See figure 6 for example, each region has an interior, boundariy, and exterior. Using these primitive , nine topological relationships can be formed by two regions, called 9-intersection model [22].

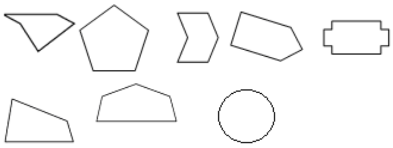


Figure 3. Example of Regions.

D. **Topological relationships**

Topology is defined as mathematical study of the properties that are preserved through all type deformations of objects. Topology is foremost a branch of mathematics, but some concepts are of importance in cartographic generalization, such as topological relationships [23]. Topological relationships describe relationships between all objects in space, the points, lines and areas for all possible kinds of deformation. Several researchers have defined topological relationships in the context of geographic information [7], [8] and [9]. In this work, we study the different transformation of topological relationships between objects when the on- the-fly map generalization was applied.

III. RELATED WORK

Historically speaking, the first algorithm for generalizing polylines was published by [19]. Then, several variants were published, essentially to improve the results of the initial algorithm. However, this algorithm does not take into account many aspects, such as the topological relationships between objects.

Then, Several methods and concepts proposed to model and implement this generalization process but a framework for their combination into a comprehensive generalization process is missing [2].

In other works, the spatial objects are model by agents, such as the works of ([3], [12], [14], [16], [17] and [18]). In this context, an important work that was suggested by [15], it present an approach based on the implementation of a multi-agent system for the generation of maps on-the-fly and the resolution of spatial conflicts. This approach is based on the use of multiple representation and cartographic generalization.

The strategy presented in [3] offers a good method for automated the generalization process; nevertheless, it is not flexible enough since it does not enable the agent to choose the best action to perform according to a given situation [15].

In the same context and for reducing the spatial conflicts in the map, a good method was proposed in [14], this method is based on the genetic algorithm. Then, the technique presented in [16] is very important, it uses the least squares adjustment theory to solve the generalization problems, but it can’t implement certain operations of generalization, such as elimination, aggregation or typification…etc.

Also, to improve the process of on-the-fly map generalization, another approach was proposed in [17] which based on a new concept called SGO (Self-generalizing object).

The work presented in [6] propose an approach based on user profiles, which formally captures the user requirements (preferences) towards the base map and deploys those profiles in a web-based architecture to generate on-demand maps.

All these methods and approaches were presented to define a good generalization process but the majority of them do not treat the transformation of topological relationships between the spatial objects when downscaling.

In this paper, we will mainly define a framework which based on the work presented in [1] and an efficient module of topological relationships that compute the topological relations between objects and propose the best transformation of them into another relations which assured the elegantly of maps. The work presented in [1] uses the multi agent system equipped with genetic algorithm in order to generate data on arbitrary scales thanks to an on-the-fly map generalization process. Thus, we will present an efficient framework combined the approach presented in [1] with the topological relationships for modelling a fast, effective and powerful generalization process.

IV. PROPOSED FRAMEWORK OF ON-THE -FLY MAP GENERALIZATION

Our study based on the work presented in [1], we use an artificial agent which can find at any time, the best solution carries out, based on its perceptions, its memory, its goals and skills. The proposed framework based on the three flowing modules:

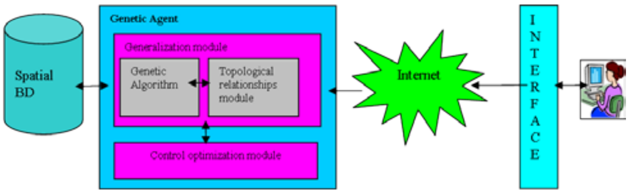


Figure 4 Framework of the on-the –fly map generalization.

A. Genetic agent Module

The roles of a genetic agent consist to generalize its self, in order to adapt it to the level of detail requested by the user. Thus, the genetic agent is responsible for the satisfaction of its constraints. It must be collaborate with the other agent, to avoid a constraints violation. The architecture of genetic agent is composed of three main modules, for more details, see our work presented in [1]:

1) Map Generalization module

This module carries out the map's generalization process; it applies the solution found by the genetic algorithm(GA). GA follows the classical steps of a genetic algorithm are selection, crossover and mutation. The solution is represented by sequence of algorithms and their good parameters. Genetic Algorithm used to find the best solutions space. In this context, we consider principally two geographical objects; ribbons and region.

For simplifying the generalization process, it can be modeled as follows [20]:

- Step 0: original geographic features only modeled as areas and/or ribbons,
- Step 1: as scale diminishes, small areas and ribbons will be generalized and possibly can coalesce,
- Step 2: as scale continues to diminish, areas mutate to points and ribbons into lines (its Skeleton),
- Step 3: as scale continues to diminish, points and lines can disappear.

Let us call this process "generalization-reduction-disappearance" (GRD process).

2) Optimization Control module

The control optimization module could achieve a satisfactory balance between discovery time of best solution and quality of the results. Thus, this module controls generalization's time for not exceeds the maximum limits and receive the message from neighboring agents which contain relevant information, such as the number of conflict agents, the distance between the neighboring objects …etc.

3) Topological relationship module

In this section, we will define a module for the transformation of topological relations during downscaling. First, we have to compute and store all the topological relations between spatial objects. Then, the mathematical assertions will be applied according to certain conditions. The main objective of this module consists to maintain the consistency of map under the geometric transformation when downscaling. Thus, this module performs the transformations of topological relationship into other ones according to mathematical assertions. These assertions based on the GRD process.

B. Topological relationships module

The generalization of spatial data implied the generalization of topological relations according to certain accurate rules. The objective of this section is to formulate the list of these rules, between regions, ribbons or regions and ribbons. The regions represent the building and ribbons represent roads or rivers. Then, sliver polygons will be taken into account in order to relax those relations, including the case of tessellations.

1) Transformation of topological Region-Region relations

In this section, the Egenhofer's relations [5] are treated mainly. After the generalization, the object geometries are adapted to the perceptual limits imposed by the new (smaller) scale. In this context, the disjoint relations transformed into meet relation. Also overlap relations transformed into cover or meet according to certain metric conditions. We use the thresholds for distance, width and areas for modeling the conditions of the assertions. We will present in this context an example of the transformation of contain relation into meet one when downscaling.

The transformation of relation "contains" into "meet", was expressed by the following assertion (Figure 5), it noted that 2Dmap is a function transforming a geographic object to some scale possibly with generalization:

$$\forall\, \mathrm{O}^1, \mathrm{O}^2 \in \mathrm{GeObject}, (\forall \sigma \in \mathrm{Scale}) \wedge (O^1_\sigma = 2Dmap(O^1,\sigma)) \wedge (O^2_\sigma = 2Dmap(O^2,\sigma)) \wedge$$
$$(Contains(\mathrm{O}^1, \mathrm{O}^2)) \wedge (Dist(\mathrm{O}^1, \mathrm{O}^2) < \varepsilon_1)$$
$$\Rightarrow Cover(O^1_\sigma, O^2_\sigma).$$

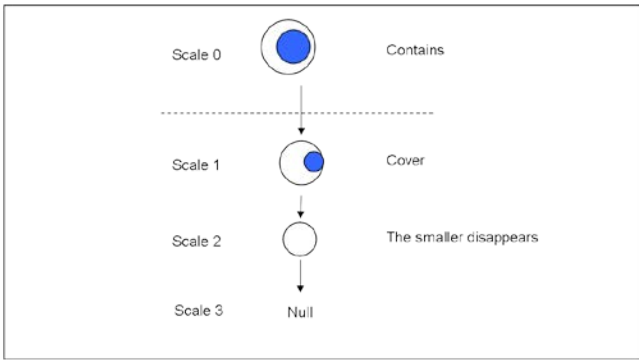


Figure 5. The transformation Contains-to-Cover

2) *Transformation of topological Ribbon-Ribbon relation*

This topological relationship between linear objects is very important because 80% of spatial objects are polyline-type [10]. In this context, we use the notion of ribbon, common examples include, disjoint, merging and crossing between ribbons. These relations represent road-road, road-river or river-river topological relationships. We take the road-road disjoint as example; this relation transformed into merging when downscaling (See Figure 6). This process can be modeled as follows:

$$\forall R^1, R^2 \in \mathrm{Ribbon}, (\forall \sigma \in \mathrm{Scale}) \wedge (R^1_\sigma = 2Dmap(R^1,\sigma)) \wedge (R^2_\sigma = 2Dmap(R^2,\sigma)) \wedge$$
$$(Disjoint(R^1,R^2)) \wedge (Dist(R^1,R^2) < \varepsilon_1)$$
$$\Rightarrow Merging(R^1_\sigma, R^2_\sigma).$$

When a Ribbon becomes very narrow, we apply this assertion:

$$\forall R \in \mathrm{Ribbon}, (\forall \sigma \in \mathrm{Scale}) \wedge (R_\sigma = 2Dmap(R,\sigma)) \wedge (Width(R_\sigma) < \varepsilon_1)$$
$$\Rightarrow R_\sigma = \phi.$$

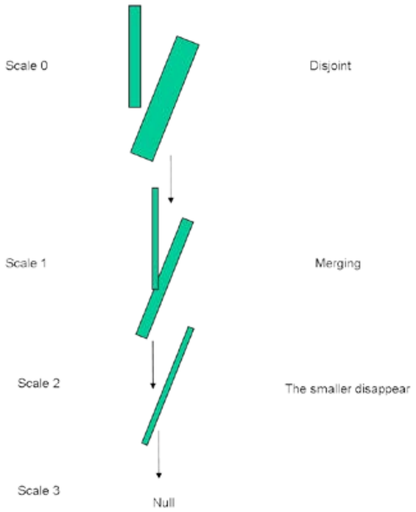


Figure 6.Transformation of disjoint relation between two ribbons.

3) *Ribbon and region*

In this section, the relations are studied which can hold between ribbon and region. To describe these relations, we based on the basic relations who may be classified into six types, namely disjoint, touches, cross, covered-by, contained-by and on-boundary, as shown in Figure 7:

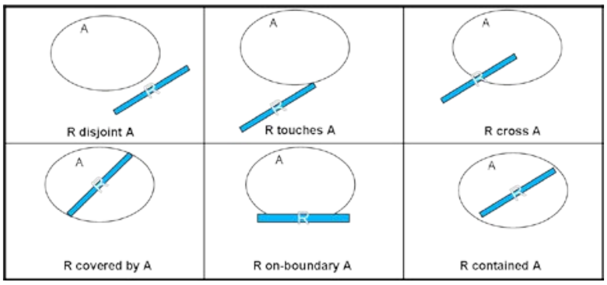


Figure 7. Basic relations between Region and Ribbon.

Therefore, one can say that any spatial relation varies according to scale. In this context, one says that a road runs along a sea; but in reality, in some place, the road does not run really along the water of the sea due to beaches, buildings, etc. At one scale, the road TOUCHes the sea, but at another scale at some places, this is a DISJOINT relation (Figure 8). Let consider two geographic objects O^1 and O^2 and $O_\sigma{}^1$ and $O_\sigma{}^2$ their cartographic representations, for instance the following assertion holds:

$$\forall\, \mathrm{O}^1, \mathrm{O}^2 \in \mathrm{GeObject}, \forall \sigma \in \mathrm{Scale} \wedge O^1_\sigma = 2Dmap(O^1,\sigma) \wedge O^2_\sigma = 2Dmap(O^2,\sigma) \wedge$$
$$Disjoint(O^1,\mathrm{O}^2) \wedge Dist(\mathrm{O}^1, \mathrm{O}^2) < \varepsilon_1 \Rightarrow Meet(O^1_\sigma, O^2_\sigma).$$

Similar assertions could be written when CONTAINS, OVERLAP relationships. In addition, two objects in the real world with a TOUCH relation can coalesce into a single one.

As a consequence, in reasoning what is true at one scale, can be wrong at another scale. So, any automatic generalization system must be robust enough to deal with this issue.

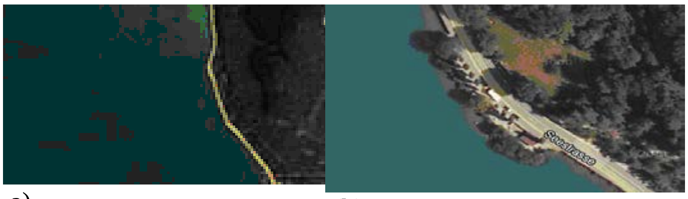


Figure 8. According to scale, the road TOUCHes or not the sea.

4) Generalized irregular tessellations when downscaling

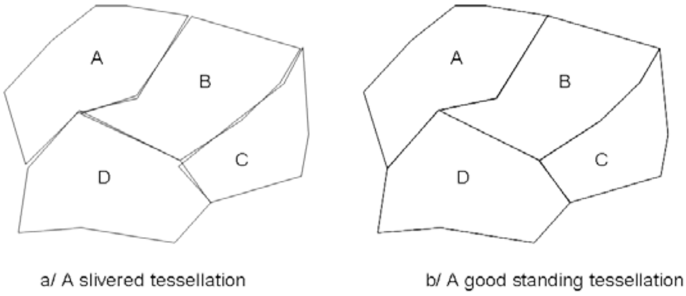
By irregular tessellation (or tessellation), one means the total coverage of an area by sub-areas. For instance the conterminous States in the USA form a tessellation to cover the whole country. Generally speaking administrative subdivisions form tessellations, sometimes as hierarchical tessellations. Let us consider a domain *D* and several polygons P_i; they form a tessellation iff (See Figure 9b):

- For any point p_k, if p_k belongs to *D* then there exists P_j, so that p_k belongs to P_j

- For any p_k belonging to P_j, then p_k belongs to *D.*

A tessellation can be also described by Egenhofer relations applied to P_i and *D*, but in practical cases, due to measurement errors, this definition must be relaxed in order to include sliver polygons (Figure 9a). Those errors are often very small, sometimes a few centimeters at scale 1. In other words, one has a tessellation from an administrative point of view, but not from a mathematical point of view.

When downscaling, those errors will be rapidly less than the threshold ε_{lp} so that the initial slivered or irregular tessellation will become a good-standing tessellation.



a/ A slivered tessellation

b/ A good standing tessellation

Figure 9. A tessellation with sliver polygons and a good standing tessellation.

The situation becomes complex when a road or river traverses this tessellation, because we have to study all topological relationships between tessellation and road or river.

V. IMPEMENTATION OF MODULE

We use in this work the Jade platform [25], JADE is a software framework, fully implemented in Java that simplifies the implementation of multi-agent systems through a middleware. JADE implements FIPA's (Foundation of Intelligent Physical Agents) specifications [24].

Using Java; in the work [1], we implement two principals' agent (building agent and road agent) and the different algorithms of generalization such as simplification, displacement and also different steps of the Genetic Agent. In this context, we will implement topological relationships module. This module computes the different relations between objects and proposes for genetic agent the best transformation of these relationships. We need to calculate the distance between objects and the area of each polygonal object. We use the Frechet distance:

A. Frechet distance

Considering two objects *A* and *B*, what is the distance between them? An interesting definition is given by the Frechet distance which corresponds to the minimum leash between a dog and its owner, the dog walking on a line, and the owner in the other line as they walk without backtracking along their respective curves from one endpoint to the other. The definition is symmetric with respect to the two curves (See Figure 10) [21]. By noting *a*, a point of *A*, and *b* of *B*, the Frechet Distance F is given as follows in which *dist* is the Euclidean conventional distance:

$$F = \underset{a\in A}{Max}(\underset{b\in B}{Min}(dist(a,b)))$$

But in our case, one must consider two distances, let us say, the minimum and the maximum of the leash, so giving:

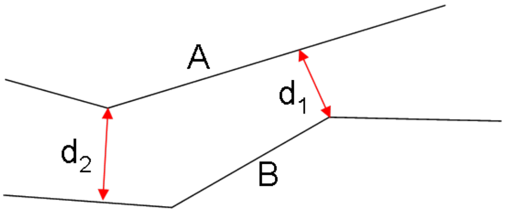


Figure 10. The Distance between two polylines.

$$d_1 = \underset{a\in A}{Min}(\underset{b\in B}{Min}(dist(a,b)))$$

and $d_2 = \underset{a\in A}{Max}(\underset{b\in B}{Min}(dist(a,b)))$.

The thresholds used in the mathematical assertions are defined from this distance. Then, the distance between two regions A and B is defined also as the Frechet distance between both boundaries. In this context, the algorithm defined in [21] is used.

We applied our framework on examples of spatial data set. These examples presented in above figures. The results are very encouraging, which show the pertinence of proposed framework.

VI. CONCLUSION

The application of the on-the-fly map generalization operators may cause topological conflicts. To avoid these conflicts, topological conditions are used to generate the relationships in terms of touching, overlapping, disjunction, and containment between map objects into others relationships.

In this context, a framework can be presented. It is used to improve the quality of the map during on-the-fly generalization. This framework based on three modules which were included in the genetic agent, as generalization module, topological relationships module and the optimization one. Thus, the genetic agent can:

- Define the best actions of generalization and it can generate its self ,
- Adapt its generalization with the other geographic agent according to scale and the topological relationships with them,
- Collaborate with the others agent to improve the result of map generalization process and resolve the spatial conflicts, in reasonable time.

This work can open various future works, such as:

- The topological relationships module did not apply all transformations of topological relations between objects. In the future, we will try to integrate the other ones.
- The mathematical assertions of the framework considered the geometries of object represented in the 2D domain; we would like to extend our work to deal with geometries of higher dimension, such as the 3D.

VII. REFERENCES

[1] B. Lejdel & O., kazar , “Genetic agent approach for improving on-the-fly web map generalization”, International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.3, June 2012.

[2] M. Bader, “Energy Minimizing Methods for Feature Displacement in Map Generalization”, Thesis (PhD), Department of Geography, University of Zurich, 2001.

[3] Duchêne C., “The CartACom model: a generalisation model for taking relational constraints into account”, 7th ICA Workshop on Generalisation and Multiple representations, Leicester, UK, 2004.

[4] Cecconi, A., “Integration of cartographic generalization and multi-scale databases for enhanced web mapping”, PhD. Thesis, University of Zurich, 2003.

[5] Egenhofer, M. “Topology and Reasoning: Reasoning about Binary Topological Relations”. In Second Symposium on Large Spatial Databases, In O. Gunther and H.-J. Schek (Eds.), LNCS in Advances in Spatial Databases, Springer-Verlag, 141-160, 1991.

[6] T. Foerster et al., “On-demand Base Maps on the Web generalized according to User Profiles”, International Journal of Geographical Information Science, Taylor & Francis, pp.1-26, 2011.

[7] Egenhofer, M., and Franzosa, R. “Point-Set Topological Spatial Relations. International Journal of Geographical Information Systems”. 5(2): 161-174, 1991.

[8] Clementini, E., Di Felice, P. , and Van Oosterom,P. “A Small Set of Formal Topological Relationships Suitable for End-User Interaction”. In Abeland, B. C. (ed), Advances in Spatial Databases, Lecture Notes in Computer Science. Springer, 692: 277-295, 1993.

[9] Winter, S., and Frank A. “Topology in Raster and Vector Representation. GeoInformatica”. 4(1):35-65, 2000.

[10] Plazanet, C., “ Enrichissement des bases de données géographiques: analyse de la géométrie des objets linéaires pour la généralisation cartographique (application au routes) ”. PhD thesis, Université de Marne-la-Vallée, 1996.

[11] Oosterom, V. and V. Schenkelaars., “The development of an interactive multi-scale GIS”, International Journal of Geographical Information Systems, Vol. 9(5):489-5071, 995.

[12] Regnauld N., “Constraints based Mechanism to achieve automatic generalization using agent modelling”, Proceedings of GIS Research UK 9th Annual Conference, 329–332, university of Glamorgan, UK, 2001.

[13] A. Newell., “Unified Theories of Cognition, Harvard University Press, Cambridge MA”, p.p 549, 1990.

[14] I.D. Wilson et al, “Reducing Graphic Conflict in Scale Reduced Maps Using a Genetic Algorithm”, 5th ICA Workshop on Progress in Automated Map Generalisation, Paris, France, 2003.

[15] N. Jabeur , “a multi-agent system for on-the-fly web map generation and spatial conflict resolution”, University of Laval, Quebec, 2006.

[16] M. Sester, “Generalization Based on Least Squares Adjustment. International Archives of Photogrammetry and Remote Sensing, vol.33, ISPRS, Amsterdam, 2000.

[17] M. Galanda, “Automated Polygon Generalization in a Multi Agent System. Dissertation zur Erlangung der

Doktorwürde, Mathematisch-naturwissenschaftliche Fakultat, Universitat Zürich, 2003.

[18] M. N. Sabo, "généralisation et des patrons géométriques pour la création des objets auto-generalisants (SGO) afin d'améliorer la généralisation cartographique à la volée ", Université de laval , Quebec, 2007.

[19] Douglas D., and Peuker T. "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature", The Canadian Cartographer. 10(2): 112-122, 1973.

[20] Laurini, R. "A Conceptual Framework for Geographic Knowledge Engineering", Journal of Visual Languages and Computing. 25(1):2-19, 2014.

[21] H. Alt, M. Godau , " Computing the Fréchet distance between two polygonal curves", International Journal of Computational Geometry and Applications, 5: 75–91, 1995.

[22] Egenhofer, M. and Herring J.,"A Mathematical Framework for the Definition of Topological Relationships", In Proceedings of the 4th International Symposium on Spatial Data Handling, 803-813, 1990.

[23] Harrie, L. "An Optimisation Approach to Cartographic Generalisation", Ph.D. Thesis, Lund University, Suède, 2001.

[24] GeoTools, " the open source java GIS toolkit", Available from: http://docs.geotools.org/, [Accessed 8 april 2014].

[25] JADE, "Java Agent Development Framework", Available from: http://jade.tilab.com/ [Accessed 8 April 2014].

A Stateless Variable Bandwidth Queuing Algorithm for Enhancing Quality of Service

Prof.C.Satheesh Pandian
Asst. Professor / CSE
Govt. College of Engg., Bargur635 104, Tamilnadu, India

The QoS (Quality of Service) is an important factor in computer networks. Here we proposed an effective queuing discipline for obtaining QoS and thus the throughput is increased. The SQP is a stateless queue management algorithm. The SQP allocates variable bandwidth for TCP (Transmission Control Protocol) flows and performs the matched drops in the queue. Based on the priority of TCP flow the queuing is done using SQP. Thus the queuing is simple and the implementation is less complex also it restricts the unresponsive TCP flows.

I. INTRODUCTION

Nowadays the network provides a connectionless, best effort and reliable data transfer. The demands on IP (Internet Protocol) based internets are rising both in terms of volume and type of service. The key design requirements for IP-based internet include congestion control, low delay, high throughput and supporting QoS.

To provide QoS the IETF is developing a suite of standards under ISA (Integrated Services Architecture) [2]. A vital element of the ISA is an effective queuing policy that takes into account the differing requirements of desired flows. A queuing policy determines which packet to transmit next if a number of packets are queued for the same output port. A separate issue is the choice and timing of packet discards. A discard policy can be an important element in managing congestion and meeting QoS.

Routers traditionally have used a first-in-first-out (FIFO) queuing discipline at each output port. When a packet arrives and is routed to an output port, it is placed at the end of the queue. In FIFO no special treatment is given for the packets from the higher priority flow. Nagle proposed a scheme called fair queuing; in this a router maintains multiple queues at each output port. A serious drawback to the fair queuing is that short packets are penalized; so a bit-by-bit approach was developed called processor sharing (PS).

The bit-round fair queuing (BRFQ) is implemented by computing virtual starting and finishing times on the fly as if PS were running. BRFQ fairly allocates the available capacity among all active flows through a node[3]. It will not provide the different amounts of the capacity to different flows with generalized processor sharing (GPS) each flow α is assigned a weight of Φ_{α}. Just as BRFQ emulates the bit-by-bit PS, WFQ emulates the bit-by-bit GPS. The strategy used by GPS is whenever a packet finishes transmission; the next packet sent is the one with the smallest value of F_i^{α} [1].

The most important example of proactive packet discard is random early detection (RED). RED [11] is designed to avoid congestion and maintain the average queue size to increase throughput and reduce average delay. In RED the average queue length *avg* is compared to two thresholds. If *avg* is less than a lower threshold TH_{min}, congestion is assumed to be minimal and the packet is placed in the queue. If *avg* is greater than or equal to an upper threshold TH_{max}, congestion is assumed to be serious and the packet is discarded. If *avg* is between the two thresholds, then onset of

ISBN: 978-1-61804-237-8

congestion is indicated. The RED will not penalize the unresponsive flows [11].

A BLUE is a queue management policy and it operates on randomly dropping or ECN marking packets in a routers queue before it overflows. A Blue queue maintains a drop/mark probability *p*, and drops/marks packets with probability *p* as they enter the queue. Whenever the queue overflows, *p* is increased by a small constant p_d, and whenever the queue is empty, *p* is decreased by a constant $p_i<p_d$.

The rest of the paper is organized as follows. Section II includes the SQP algorithm and section III includes the performance results. We conclude our paper in section IV.

II. SQP ALGORITHM

SQP uses the strategy of matched drops presented by CHOKe [19] to protect TCP flows. Like CHOKe, SQP is a stateless algorithm that is capable of working in core networks where a myriad of flows are served. More importantly, SQP supports differentiated bandwidth allocation for traffic with different priority weights. Each priority weight corresponds to one of the priority levels; a heavier priority weight represents a higher priority level.

Although SQP borrows the idea of matched drops from CHOKe for TCP protection, there are significant differences between these two algorithms. First of all, the goal of CHOKe is to block high-speed unresponsive flows with the help of RED to inform TCP flows of network congestion, whereas SQP is designed for supporting differentiated bandwidth allocation with the assistance of matched drops that are also able to protect TCP flows.

While the authors of [19] suggested to draw more than one packet if there are multiple unresponsive flows, they did not provide further solutions. In SQP, the adjustable number of draws is not only used for restricting the bandwidth share of high-speed unresponsive flows, but also used as signals to inform TCP of the congestion status. In order to avoid functional redundancy, SQP is not combined with RED since RED is also designed to inform TCP of congestion. Thus we say that SQP is an independent AQM scheme, instead of an enhancement filter for RED. Simulations were used to compare the performance of SQP and that of SQPRED (i.e., SQP with RED) and the results are similar. We do not show the simulation results in this paper due to the space limit.

In order to determine when to draw a packet (or packets) and how many packets are possibly drawn from the buffer, we introduce a variable, called the drawing factor, to control the time as well as the maximum number of draws. For flow I (i = 1, 2, ... , N, where N is the number of active flows), the drawing factor is denoted by pi (pi > 0). Roughly speaking, we may interpret pi as the maximum number of random draws from the buffer upon an arrival from flow i. The precise meaning is discussed below. Let wi (wi > 1) denote the priority weight of flow i. If two flows, say, i and j, are given the same priority, then wi = wj. Let p0 denote the drawing factor for traffic with the lowest

priority. The relationship among wi, pi, and po can be described as

$$Pi = Po/wi. \qquad (1)$$

If flow i has a higher priority weight than flow j (wi > wj), flow i will have a smaller drawing factor than flow j (Pi < pj). In other words, the maximum number of random draws upon an arrival from flow i is smaller, and hence flow i will have a lower possibility of becoming the victim of matched drops. This is the basic mechanism for supporting bandwidth differentiation in SQP. In addition, from (1) we have Pi = pj when flow i and flow j are of the same priority; this is how SQP provides better fairness among flows with the same priority). Now we discuss the precise meaning of drawing factor Pi, which depends upon its value. According to the value of Pi (pi > 0), the drawing process can be categorized into two cases:

Case 1. When 0 < pi < 1, pi represents the probability of drawing one packet from the buffer at random for comparison.

Case 2. When Pi > 1, Pi consists of two parts, and we may rewrite

Pi as pi = m+ f. (2)

where m C E* (the set of nonnegative integers) represents the integral part with the value of [pi] (the largest integer < Pi), and f represents the fractional part of Pi. In this case, at the most m or m + 1 packet in the buffer may be drawn for comparison. Let d_{max} denote the maximum number of random draws. We have

$\text{Prob}[d_{max}= m+1]\ f,$

$\text{Prob}[d_{max}= m] = 1-f.$

```
For each packet (pkt) arrival
(1) Update $p_o$
(2) Check the priority level of pkt
     IF it corresponds to weight $w_i$
    THEN $pi \leftarrow po/wi, m \leftarrow [pi], f = pi-m$
(3) Generate a random no v €(0,1)
     IF $v < f$
     THEN $m \leftarrow m +1$
(4) IF $L>L_{th}$
    THEN
        WHILE m > 0
            $m \leftarrow m -1$
            Draw a packet (pkt') at random
            IF pkt' and pkt are from the same flow
            THEN
                Drop both pkt' and pkt
                Return
            ELSE keep pkt' intact
    IF buffer is full
    THEN drop pkt
    ELSE let pkt enter the buffer
```

Parameters:

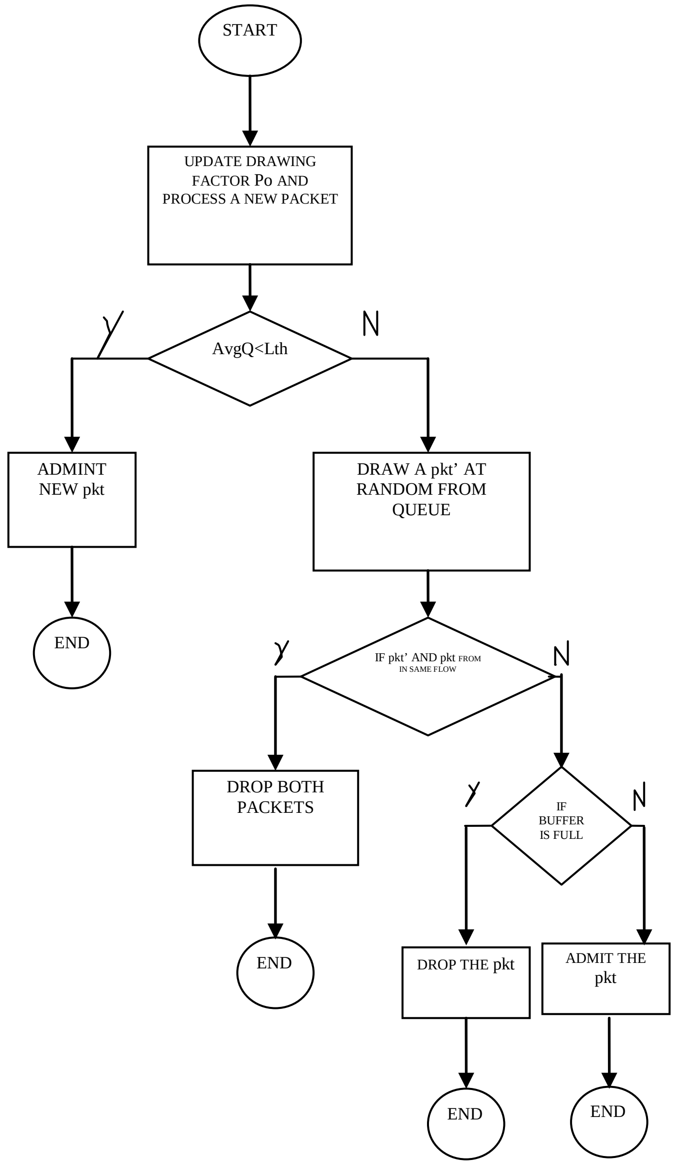
L **: Queue length**

L_{th} **: Queue length threshold to activate drawing**

Fig. 1. The algorithm of drawing packets

The algorithm of drawing packets is described in Fig. 1. In this figure, we only use one variable m to record the value of m (before Step (3)) and d_{max} (after Step (3)). Note that no more packets need to be drawn as soon as matched drops occur. On the other hand, if no match is found, the drawing process will continue until the number of draws reaches d_{max}.

A. Flow Diagram


START
UPDATE DRAWING FACTOR Po AND PROCESS A NEW PACKET
AvgQ<Lth
Y
N
ADMINT NEW pkt
DRAW A pkt' AT RANDOM FROM QUEUE
END
IF pkt' AND pkt FROM IN SAME FLOW
Y
N
DROP BOTH PACKETS
IF BUFFER IS FULL
Y
N
END
DROP THE pkt
ADMIT THE pkt
END
END

```
Initialization:
$PO \leftarrow 0$
IF $L < L^-$
THEN
  $po \leftarrow Po-P^-$
  IF $po < 0$
  THEN $po \leftarrow 0$
IF $L > L^+$
```

THEN $po \leftarrow po + p^+$

Parameters:
L**: Queue length**
L^+ **: Queue length threshold to increase** po
L^- **: Queue length threshold to decrease** po
$L_{th} < L^- < L^+$
po**: Basic drawing factor with initial value 0**
$p+$**: Step length required for increasing** po
$p-$**: Step length required for decreasing** po

Fig.2.The algorithm of updating ***p0***

This adaptive function is implemented by updating p0. The updating process is shown in Fig.2, which details Step (1) of the algorithm of drawing packets that is shown in Fig. 1. The combination of Fig. 1 and Fig.2 provides a complete description of the SQP algorithm. The present state of SQP can be described by the activation of matched drops and the process of updating po, which is further determined by the range the current queue length L falls into, shown in Table I. At any time, SQP works in one of following states:

1) inactive matched drops and decreasing *po* (unless *p0* = 0), when $0 < L < L_{th}$;

2) active matched drops and decreasing po (unless *po* = 0), when $Lth < L < L^-$;

3) active matched drops and constant po, when $L^- < L < L^+$;

4) active matched drops and increasing *po*, when $L+ < L < L_{lim}$.

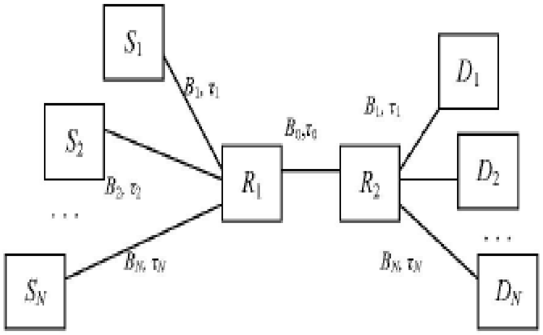


Fig. 3. Network Topology

III. PERFORMANCE EVOLUTION

To evaluate SQP in different scenarios and to compare it with some other schemes, we implemented SQP using ns simulator version 2 [17]. The SQP code is designed as a patch for ns, and it is available at [26]. In this section, we use the network topology shown in Fig.3 where Bo 1 Mb/s and Bi = 10 Mb/s (i = 1, 2,... , N). Unless specified otherwise, the link propagation delay To = Ti 1 Ims. The buffer limit is 500 packets, and the mean packet size is 1000 bytes. TCP flows are driven by FTP applications, and UDP flows are driven by CBR traffic. All TCPs are simulated as TCP SACK. Each simulation runs for 500 seconds.

Parameters of SQP are set as follows: Lth = 100 packets, L- = 125 packets, L+ = 175 packets, p+ = 0.002, and p- = 0.001.

Parameters of RED are set as follows: minth = 100 packets, maxth = 200 packets, gentle = true, pmax = 0.02, and the EWMA weight is set to 0.002.

Parameters of RIO include those for "out" traffic and those for "in" traffic. For "out" traffic, min_{th} out = 100 packets,

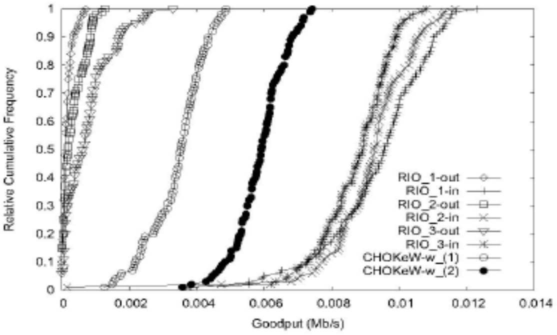


Fig. 4. The Relative Cumulative Frequency of RIO and CHOKeW

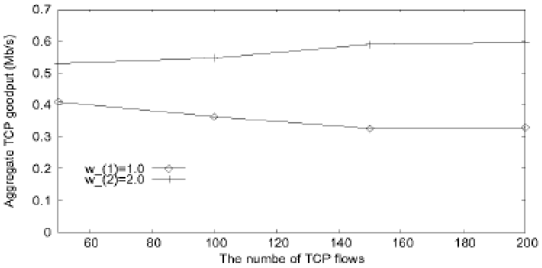


Fig. 5. The aggregate TCP goodput for two priority levels

Goodput (Mb/s) 0.01 0.012 0.014 Fig. 4. The Relative Cumulative Frequency of RIO and SQP max_{th} out = 200 packets, p_{max} out = 0.02. For "in" traffic, min_{th} in = 110 packets, maxth in = 210 packets, p_{max} in = 0.01. Both gentle out and gentle in are set to true.

Parameters of BLUE are set as follows: 1 = 0.0025 (the step length for increasing the dropping probability), 62 = 0.00025 (the step length for decreasing the dropping probability), and freeze time= 100 ms. From Fig.4 we see that for "out" traffic of RIO 1 the RCF of goodput zero is 0.1. In other words, 10 of the 100 "out" flows are starved. Similarly, for RIO 2 and RIO 3, 15 and 6 flows are starved, respectively. Even some "in" flows of RIO may have very low goodput (e.g., the lowest goodput of "in" flows of RIO2 is only 0.00015 Mb/s). Flow starvation is very common in RIO, but it is rarely observed in SQP.

Now we investigate the relationship between the number of TCP flows and the aggregate TCP goodput for each priority level. The results are shown in Fig.5. Here two priority levels are corresponding to w(i) = 1 and W(2) = 2. Half of the flows are assigned w(i) and the other half assigned W(2).

IV. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a cost effective active queue management called SQP that process various flows in different priority levels. Both the analytical model and simulations of SQP are shown. Further work involves studying the algorithm with different parameters and network topologies.

REFERENCES

[1] J. Bennet and H. Zhang, WF2Q: Worst Case Fair Weighted Fair Queuing,
IEEE INFOCOM'96, 1996

[2] S. Blake, D. Black, M. Carlson, et al., An Architecture for Differentiated Service, IETF RFC2475,December 1998

[3] U. Bodin, 0. Schelen, and S. Pink, Load-Tolerant Differentiation with Active Queue Management, ACM CCR'00. [Online]. Available: http://www.acm.org/sigcomm/ccr/archive/ccr-toc/ccr-toc-2000.html [4] R. Braden, D. Clark, and J. Crowcroft, et al., Recommendations on Queue Management and Congestion Avoidance in the Internet, IETF RFC 2309, April 1998

[5] K. Cho, Flow-Valve: Embedding a Safety-Valve in RED, IEEE GLOBECOM' 99

[6] R. B. Cooper, Introduction to Queueing Theory, 2nd ed. Elsevier North Holland, 1981

[7] D. D. Clark, S. Shenker, and L. Zhang, Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanism, ACM SIGCOMM'92

[8] D. D. Clark and W. Fang, Explicit Allocation of Best Effort Packet Delivery Service, IEEE/ACM Transactions on Networking, August 1998, 6(4):362-373

[9] A. Demers, S. Keshav, and S. Shenker, Analysis and simulations of a Fair Queueing algorithm, ACM SIGCOMM'89

[10] S. Floyd and V. Jacobson, Random Early Detection Gateways for Congestion Avoidance, IEEE/ACM Transaction on Networking, Aug.1993, 1(4):397-413

[11] S. Floyd, RED: Discussions of Setting Parameters, November 1997. [Online]. Available: http://www.icir.org/floyd/REDparameters.txt

[12] S. Floyd and K. Fall, Promoting the Use of End-to-End Congestion Control in the Internet, IEEE/ACM Transaction on Networking,
Aug.1999, 7(4):458-472

[13] J. Heinanen, F. Baker, W. Weiss, et al., Assured Forwarding PHB Group, IETF, RFC 2597, 1999

[14] R. Mahajan and S. Floyd, Controlling High-Bandwidth Flows at the Congested Router, ICSI Tech Report TR-01-001, April 2001. [Online]. Available: http://www.icir.org/red-pd/

[15] P. Marbach, Pricing Differentiated Services Networks: Bursty Traffic, IEEE INFOCOM'01

[16] N. Nichols, S. Blake, F. Baker, et al., Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers, IETF RFC 2474, Dec. 1998

[17] ns-2 (Network Simulator version 2). [Online].
vailable:http://www.isi.edu/nsnam/ns/

[18] J. Padhye, V. Firoiu, D. Towsley, et al., Modeling TCP Throughput: A Simple Model and its Empirical Validation, ACM SIGCOMM'98

[19] R. Pan, B. Prabhakar, and K. Psounis, CHOKe: A Stateless Active Queue Management Scheme for Approximating Fair Bandwidth Allocation, IEEE INFOCOM'01

[20] A. K. Parekh and R. G. Gallager, A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: the Single- Node Case, IEEE/ACM Transaction on Networking, Aug. 1993, 1(3):344- 357

[21] A. K. Parekh and R. G. Gallager, A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: the Multiple- Node Case, IEEE/ACM Transaction on Networking, Aug. 1994, 2(2): 137- 150

[22] S. Ramabhadran and J. Pasquale, Stratified Round Robin: A Low Complexity Packet Scheduler with Bandwidth Fairness and Bounded Delay,
ACM SIGCOMM'03

[23] I. Stoica, S. Shenker, and H. Zhang, Core-Stateless Fair Queueing: Achieving Approximately Fair Bandwidth Allocation in High Speed Networks, ACM SIGCOMM'98

[24] S. Suri, G. Varghese, and G. Chandramenon, Leap Forward Virtual Clock: A New Fair Queueing Scheme with Guaranteed Delay and Throughput Fairness, IEEE INFOCOM '97, April 1997

[25] A. Tang, J. Wang, and S. H. Low, Understanding CHOKe, IEEE INFORCOM'03

[26] S. Wen and Y. Fang, CHOKeW Patch on ns, April 2005. [Online]. Available: http://www.ecel.ufl.edu/ wen/chokew.zip

Authors Index