ADVANCES in INFORMATION SCIENCE and APPLICATIONS - VOLUME I

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 I

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 | 22

ISSN: 1790-5109
ISBN: 978-1-61804-236-1

ADVANCES in INFORMATION SCIENCE and APPLICATIONS - VOLUME I

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

ISBN: 978-1-61804-236-1

ISBN: 978-1-61804-236-1

ISBN: 978-1-61804-236-1

ISBN: 978-1-61804-236-1

ISBN: 978-1-61804-236-1

ISBN: 978-1-61804-236-1

ISBN: 978-1-61804-236-1

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

ISBN: 978-1-61804-236-1

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).

ISBN: 978-1-61804-236-1

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,

ISBN: 978-1-61804-236-1

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.

ISBN: 978-1-61804-236-1

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.

ISBN: 978-1-61804-236-1

PART I

ISBN: 978-1-61804-236-1

ISBN: 978-1-61804-236-1

A Comparative Analysis of Binary Patterns with Discrete Cosine Transform for Gender Classification

Marcos A Rodrigues, Mariza Kormann and Peter Tomek

***Abstract*—**This paper presents a comparative analysis of binary patterns for gender classification with a novel method of feature transformation for improved accuracy rates. The main requirements of our application are speed and accuracy. We investigate a combination of local binary patterns (LBP), Census Transform (CT) and Modified Census Transform (MCT) applied over the full, top and bottom halves of the face. Gender classification is performed using support vector machines (SVM). A main focus of the investigation is to determine whether or not a 1D discrete cosine transform (DCT) applied directly to the grey level histograms would improve accuracy. We used a public database of faces and run face and eye detection algorithms allowing automatic cropping and normalisation of the images. A set of 120 tests over the entire database demonstrate that the proposed 1D discrete cosine transform improves accuracy in all test cases with small standard deviations. It is shown that using basic versions of the algorithms, LBP is marginally superior to both CT and MCT and agrees with results in the literature for higher accuracy on male subjects. However, a significant result of our investigation is that, by applying a 1D-DCT this bias is removed and an equivalent error rate is achieved for both genders. Furthermore, it is demonstrated that DCT improves overall accuracy and renders CT a superior performance compared to LBP in all cases considered.

***Keywords*—**Image processing, feature extraction, gray-scale, image texture analysis, pattern recognition, discrete cosine transforms, support vector machines

I. Introduction

REAL-time gender classification is a requirement for marketing applications where legal and ethical constraints do not allow the saving of images either locally or remotely for later processing. The ADMOS project [1] is funded by the European Union and aims to develop a real-time gender classification and age estimation to be used in private spaces of public use, such as shopping malls, fairs and outdoor events. The main computing operations on an image within the time frame of live capture include face detection, gender classification and age estimation. We are investigating a number of methods that have the potential to be fast, accurate and robust. In this paper we report on a combination of techniques involving LBP–Local Binary Patterns, CT–Census Transform, MCT–Modified Census Transform, DCT–Discrete Cosine Transform and SVM–Support Vector Machines. It is shown that DCT can remove LBP's bias towards higher accuracy for male subjects and that it renders CT a superior technique when compared to LBP.

LBP is a non-parametric method used to summarise local structures of an image and have been extensively exploited in face analysis for gender, age, and face recognition [2], [3], [4], [5], [6], [7]. Normally, LBP are employed in local and holistic approaches and a number of extensions have been demonstrated in the literature (e.g. [4]) in connection with linear discriminant analysis and support vector machines. The Census Transform is similar to LBP; the main difference lies on how bits are concatenated together. Although this is a seemingly small difference, it has significant bearings on the final grey level scale histograms and thus, on the texture descriptors in various regions of an image.

The Census Transform has not been extensively exploited in face analysis as LBP; some previous work include [8], [9]. LBP and DCT have been used together in connection with face recognition and gender classification (e.g. [10], [11], [12]). However, it is important to note that when DCT is used, it is invariably in connection with a 2D-DCT. Normally a DCT is performed over the entire input image using different block sizes. In a similar fashion, LBP is normally applied over regions and over the entire image and such histograms are concatenated into a combined one.

Here we explore these techniques aiming at fast processing for real time applications. We only use a single pass, non-optimised LBP, CT or MCT over the input image, followed by a 1D-DCT applied to the resulting histograms. The purpose is to investigate whether or not the 1D-DCT would improve gender classification. The approach is demonstrated by using a public database from which the various regions of interest are automatically selected by face and eye detection algorithms. It is shown that 1D-DCT improves gender classification in all cases considered.

The method is described in Section II, experimental results are presented in Section III with conclusions and further work in Section IV.

II. Method

The approach to gender classification adopted in this paper has been described in our previous paper [13] summarised as follows:

- Define a set of measurements or features $m_i (i = 1, 2, \ldots, N_1)$ over an input image and build a vector $M_j = (m_1, m_2, \ldots, m_{N_1})^T$, with $j = 1, 2, \ldots, N_2$ characterising the selected features;
- Build a matrix Ω_k of vectors M_j where the index of k points to the identity of the input vector: $\Omega_k = (M_1, M_2, \ldots, M_s)^T$ where s is the total number of vectors for class k;
- Define a method to estimate the closest distance to a given vector M to the most similar vector in the database. The class of closest vector(s) will point to the most likely class of M.

Arguably the most critical step is feature selection. In [13] we used LBP in connection with Eigenvector decomposition to determine class membership. Only raw data were used with no

This project has received funding from the European Union Seventh Framework Programme for research, technological development and demonstration under grant agreement number 315525, 2013–2015.

Marcos A Rodrigues and Mariza Kormann are with the GMPR–Geometric Modelling and Pattern Recognition Research Group at Sheffield Hallam University, Sheffield, UK. Email {*m.rodrigues, m.kormann*}@shu.ac.uk

Peter Tomek is with ATEKNEA, Budapest, Hungary. Email peter.tomek@ateknea.com

ISBN: 978-1-61804-236-1

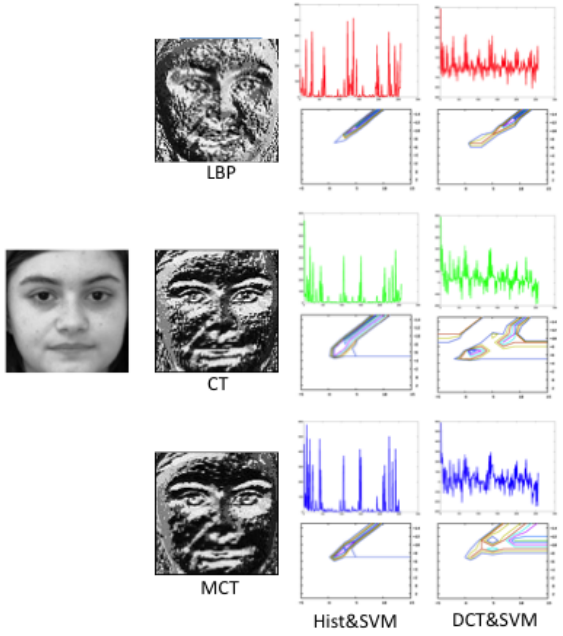


Fig. 1. The proposed method: histograms from LBP, CT and MCT are transformed by DCT. Both original histograms and their DCT are compared through SVM.

feature optimisation. The purpose of that study was to determine which ROI–region of interest would be more appropriate for robust gender classification. Tests were reported on using a larger image comprising the head with portions of the neck; this invariably also included large chunks of background and, as expected, did not yield robust results. Further experiments on cropped regions of the face used the full, top and bottom halves of the face and various combinations of these. It was shown that, for the dataset used, the best region was the top half of the face. Gender classification accuracy of 88% was reported on non-optimised data and non-optimised classification method. It was pointed out that other feature selection techniques could be used and that a literature review pointed to SVM as a robust technique in connection with LBP.

In this paper we propose a new method for gender classification with a comparative analysis of performance. The method uses the LBP, Census Transform and Modified Census Transform for feature extraction, discrete cosine transform for feature transformation and support vector machines for classification. In particular we are interested in determining whether or not a similar technique to LBP, the Census Transform and its modified version would yield better, worst or indifferent results. Furthermore, whether or not the discrete cosine transform can be effectively used for gender classification in connection with such feature extraction techniques. The steps in the proposed method are depicted in Figure 1 described as follows:

1) Apply LBP on input images and build training and test sets for both female and male subjects. The size of the kernel window is 3×3;
2) Apply the Census Transform to the same images and build training and test sets for both classes, with kernel window size of 3×3;
3) Apply a Modified Census Transform to the same images building training and test sets for both classes, with kernel window size 3×3;
4) Apply the discrete cosine transform to the outputs from steps 1–3;
5) Use SVM for training and testing all data from steps 1–4.

The proportion of data that is used for training and testing can vary considerably; in this paper we use 70–30 (70% of all data for training and 30% for testing).

A. The Census Transform – CT

The Census Transform has been proposed in [14] as a grey-level operator over a local neighbourhood. It applies to an image kernel of size $m \times n$:

$$CT_{m,n}(x,y) = \Big\|_{i=-n/2}^{n/2} \Big\|_{j=-m/2}^{m/2} f(I(x,y), I(x+i, y+j)) \quad (1)$$

where the operator $||$ is a bit-wise concatenation of $f(u,v)$ which is defined as

$$f(u,v) = \begin{cases} 0 & \text{if } u \leq v, \\ 1 & \text{otherwise.} \end{cases}$$

Various modifications have been suggested to the original CT transform such as centre-symmetric weighted kernel [15] and a modified CT using the mean of the centre pixel [16]. Typical window sizes are 3×3, 5×5 and 9×7 as their concatenated binary results fit into 8, 32 or 64 bit. Experiments have shown (e.g. [17]) that using a kernel window of 5×5 is a good compromise between speed and accuracy.

The Modified Census Transform (MCT) as used in this paper is similarly defined as in equation 1. The difference is that instead of using the grey level intensity of the centre pixel, the average of the kernel window intensity is used.

B. Local Binary Patterns – LBP

Local binary patterns [2], [18], [19], [20] are grey-scale operators defined over local neighbourhood pixels. It was originally defined using a 3×3 array of pixels, but many implementations consider larger radii. The value of the centre pixel is compared with its neighbours and the result (greater or smaller) expressed as a binary number and concatenated over all pixels considered. The concatenated array of binary numbers is normally converted to grey scale from which histograms are produced. LBP can be expressed over P sampling points on a circle of radius R where the value of the centre pixel (x, y) is expressed as:

$$\text{LBP}_{P,R} = \sum_{p=0}^{P-1} T(I_p - I_c) 2^p, \quad (2)$$

where $I_p - I_c$ is the difference of pixel intensity in the grey level between a current pixel and centre pixel of the kernel window. P is the number of pixels on a circle of radius R, and T is a thresholding function defined as:

$$T(.) = \begin{cases} 1 & \text{if } (I_p - I_c) \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

In order to improve the discriminating power of LBPs, images are normally defined in blocks from which individual LBPs are calculated and then concatenated into a single histogram. The analysis of such histograms can be used to differentiate texture patterns. A number of variants to LBP have been proposed in the literature (e.g. [3], [4], [5]). In the experiments reported in this paper we only consider the original LBP definition.

ISBN: 978-1-61804-236-1

C. The Discrete Cosine Transform – DCT

The DCT transform and its variants have been used in a number of contexts most notably in image and video compression (e.g. [21], [22], [23]). DCT is a close relative to the discrete Fourier transform as it defines a sequence of data in terms of the sum of the cosine functions at different frequencies. There are many versions of the DCT and here we use the unitary discrete cosine transform as defined in Matlab [24]. The DCT transform of a one-dimensional signal z (in our case z is an image histogram) is expressed as:

$$y(k) = w(k) \sum_{n=1}^{N} z(n) \cos(\frac{\pi(2n-1)(k-1)}{2N}) \qquad (3)$$

for $k = 1, 2, \ldots N$ where N is the length of the signal and

$$w(k) = \begin{cases} 1/\sqrt{N} & \text{for } k = 1, \\ \sqrt{2/N} & \text{for } 2 \leq k \leq N. \end{cases} \qquad (4)$$

The length of the coefficients y is the same as the original signal z. A useful property of the DCT is that normally it is only necessary a few coefficients to reconstruct the signal; most signals can be described with over 99% accuracy by using only a handful of coefficients. Here we choose to use all coefficients for improved accuracy.

D. Support Vector Machines – SVM

In pattern recognition tasks, algorithms for linear discriminant functions can be used either over the raw or original data features or in a transformed space that can be defined by nonlinear transformations of the original variables (e.g. DCT applied to the LBP and CT histograms as proposed in this paper). Support vector machines are algorithms that implement a mapping of pattern vectors to a higher dimensional feature space and find a 'best' separating hyperplane between the data set. The best hyperplane is commonly referred to as the maximal margin hyperplane, as it is defined where the closest points between classes are at maximum distance [25].

Given a set of M training samples $(l_i, \mathbf{x}_i)$ where l_i is the associated class label ($l_i \in \{-1, 1\}$) of vector $\mathbf{x}_i$ where $\mathbf{x}_i \in R^N$, a SVM classifier finds the optimal hyperplane that maximises the margin between classes l_i:

$$f(x) = \sum_{i=1}^{M} l_i \alpha_i \cdot k(\mathbf{x}, \mathbf{x}_i) + b \qquad (5)$$

where $k(\mathbf{x}, \mathbf{x}_i)$ is a kernel function, b is a bias and the sign of $f(x)$ is used to determine the class membership of vector $\mathbf{x}$. For a two-class problem (e.g. the case of gender classification) a linear SVM might suffice. In this case, the kernel function is a dot product in the input space.

III. Experimental results

A public database as described in [13] is used here, details of the database can be found in [26]. It contains 2,779 images of even balanced number of male and female subjects captured with large variations in pose and illumination. In [13] a set of 50 male and 50 female subjects were used; although all subjects were not in strictly frontal pose, there was not much variation in illumination in that dataset. Here we expand the dataset by including images with uneven illumination. We used the entire set of 99 male and 99 female subjects (one subject was removed from each original set of 100 as they appear to be repeated). The selected data allows the testing of algorithms in a realistic scenario of uneven illumination. Examples of selected images from the database are depicted in Figure 2.

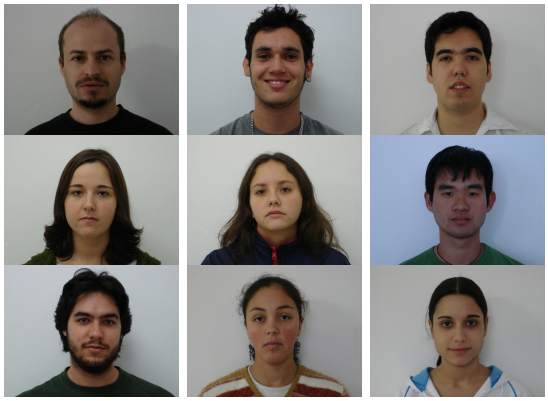


Fig. 2. Examples of image data from the FEI database.

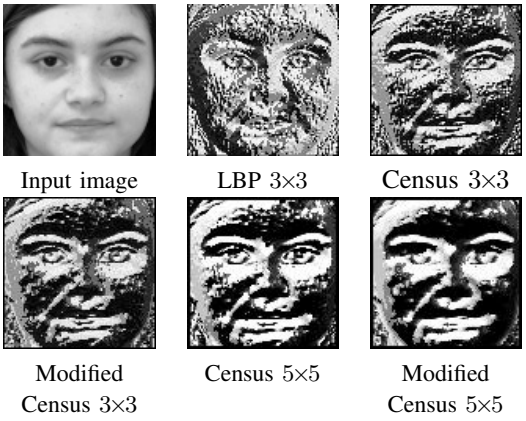


Fig. 3. LBP, Census and Modified Census over an input image.

First, we performed a visual comparison of LBP, CT and MCT using a 3×3 kernel window. We proceeded to perform a CT and MCT using a 5×5 window as shown in Figure 3. It is not clear whether or not simply increasing the kernel size would impact on gender classification performance and this is left for further studies. Here we report on LBP, CT and MCT with constant 3×3 kernel window.

Following results described in [13] we use the face regions labeled as FULL face, TOP and BOTTOM half of the face. Histograms for LBP, CT and MCT were evaluated as described in Section II. From these we also built concatenated histograms for FULL||TOP, FULL||BOTTOM, and TOP||BOTTOM. Furthermore, we also built concatenated histograms of LBP||CT and LBP||MCT for all cases. The histograms were then separated into training and test sets and subject to the SVM discrimination method.

Results for histogram-based classification are tabulated in Table I. The summary refers to training on 60 data sets (30 female, 30 male) and 60 test sets. The best result is for LBP applied to the TOP half of the face, and this confirms previous results as reported in [13]. Overall, the LBP technique is shown to be superior to CT and MCT and the concatenated combinations with the overall lowest standard deviation of 6.4. Furthermore, it is observed that there is a bias towards more accurate male classification as reported in all papers in the literature; this is the case for all combinations used. The reasons for this behaviour are not yet clear.

Following this initial comparison, all histograms were subject to discrete cosine transform, trained and tested with the

ISBN: 978-1-61804-236-1

TABLE I
ACCURACY OF HISTOGRAM-BASED CLASSIFICATION (%)

Face ROI & Method	**LBP**	**CT**	**MCT**	**LBP|| CT**	**LBP|| MCT**
FULL					
Female	80.6	87.1	67.7	83.9	77.4
Male	90.3	93.5	83.9	90.3	83.9
TOP					
Female	83.9	87.1	74.7	83.9	80.6
Male	96.8	87.1	77.4	90.3	90.3
BOTTOM					
Female	80.6	74.2	67.7	71.0	74.2
Male	93.5	93.5	93.5	96.8	93.5
FULL||TOP					
Female	83.4	87.1	77.4	77.4	87.1
Male	93.5	93.5	87.1	96.8	87.1
FULL||BOTTOM					
Female	83.9	87.1	64.5	71.0	77.4
Male	90.3	93.5	90.3	96.8	96.8
TOP||BOTTOM					
Female	80.6	67.7	74.2	83.9	80.6
Male	96.8	100.0	90.3	93.5	93.5
Mean	87.9	87.6	79.0	86.3	85.2
STD	6.4	8.9	9.9	9.4	7.3
Mean Female	82.2	81.7	71.0	78.5	79.6
Mean Male	93.5	93.5	87.1	94.1	90.9
Abs difference	11.4	11.8	16.1	15.6	11.3

TABLE II
ACCURACY OF DCT-BASED CLASSIFICATION (%)

Face ROI & Method	**LBP**	**CT**	**MCT**	**LBP|| CT**	**LBP|| MCT**
FULL					
Female	93.5	93.5	80.6	93.5	93.5
Male	87.1	90.3	83.9	87.1	87.1
TOP					
Female	90.3	90.3	71.0	93.5	80.6
Male	83.9	90.3	77.4	90.3	83.9
BOTTOM					
Female	80.6	80.6	71.0	80.6	74.2
Male	77.4	87.1	77.4	77.4	80.6
FULL||TOP					
Female	93.5	93.5	93.5	90.3	93.5
Male	87.1	90.3	80.6	87.1	87.1
FULL||BOTTOM					
Female	90.3	93.5	87.1	90.3	90.3
Male	83.9	87.1	74.1	83.9	83.9
TOP||BOTTOM					
Female	87.1	87.1	93.5	87.1	90.3
Male	87.1	93.5	83.9	87.1	83.9
Mean	86.8	89.8	81.2	87.4	85.7
STD	4.8	3.8	7.6	4.8	5.7
Mean Female	89.2	89.9	82.8	89.2	87.1
Mean Male	84.4	89.8	79.6	85.5	84.4
Abs difference	4.8	0.0	3.2	3.7	2.6

SVM method as per previous sets. Results are tabulated in Table II. Two important observations can be made: the overall accuracy has improved for all sets and the bias towards higher male accuracy has been removed. Furthermore, the CT is now shown to be a superior technique with the lowest standard deviation of 3.8. With the removal of bias, both female and male classification are equivalent, with absolute difference between their means ("Abs difference") ranging from 0.0–4.8. This favourably compares to 11.3–16.1 of previous set of experiments. These results demonstrate that the discrete cosine transform can effectively be applied over the grey level histograms for improved gender classification.

IV. CONCLUSIONS

This paper has presented a comparative analysis of gender classification based on LBP, CT and MCT in connection with the discrete cosine transform and support vector machines. The new proposed method is based on evaluating binary patterns and building histograms of various regions of the face including the full face, top and bottom halves, and a combination of these. Histograms are then subject to the discrete cosine transform. Discriminant analysis is performed through support vector machines both on the original and DCT-transformed histograms.

If only histograms are used (including various concatenations as reported here) it is shown that LBP is a more accurate technique than either CT and MCT (LBP is only marginally more accurate than CT, but it has a much lower overall standard deviation making it a more robust technique). It is also observed that the best classification results are obtained over the top half of the face, a region that includes the front and the eyes. This confirms previous results from our research reported in [13]. Furthermore, it is noted that there is a bias towards more accurate classification over male subjects, as reported in the literature. Some explanations for why this is so has been attempted in the literature but it is pointed out (e.g. [15]) that a thorough analysis is required to explain this behaviour.

Experimental results for DCT-transformed histograms demonstrate that the CT method is the more accurate technique. A number of observations can be made: overall accuracy is improved for all cases, the standard deviation is substantially decreased for all cases, and the bias towards higher accuracy for male subjects is removed. We do not yet offer an explanation for this as a detailed mathematical analysis is required, which is left for further studies.

The results reported in this paper clearly show that the discrete cosine transform yields more appropriate data for accurate classification. Further work includes applying the transformations on histograms of concatenated data from sub-regions of the face – the principle that has been shown in the literature is that it seems that the more features are used the better the accuracy. Obviously that there is a limit to the number of features and this is an area for further investigation. Moreover, we intend to apply the techniques described here on other public databases namely FERET and SUMS. It appears that a larger number of different algorithms have been applied to these databases as reported in the literature and this will provide a more direct comparison of performance.

REFERENCES

[1] ADMOS (2013). *Advertising Monitoring System Development for Outdoor Media Analytics*, EC Grant Agreement 31552. [Online] Available at http://admos.eu

[2] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen (2011). *Computer Vision Using Local Binary Patterns*. Springer.

[3] J, Ylioinas, A. Hadid, M. Pietikäinen (2011). Combining contrast and local binary patterns for gender classification. *SCIA 17th Scandinavian Conference on Image Analysis*, 2011.

[4] C. Shan (2012). Learning local binary patterns for gender classification on real-world face images, *Pattern Recognition Letters* 33 (2012) 431–437.

[5] Y. Guo, G. Zhao, M. Pietikäinen, and Z. Xu (2010). Descriptor learning based on fisher separation criterion for texture classification. *In Proc. ACCV2010*, 185–198, 2010.

[6] H. Lian, B. Lu (2007). Multi-view gender classification using multi-resolution local binary patterns and support vector machines. *Int J Neural Systems* 17 (6), 479–487.

[7] N. Sun, W. Zheng, C. Sun, C. Zou, L. Zhao (2006). Gender classification based on boosting local binary pattern. *In: Int Symp on Neural Networks*, 2006.

[8] R. Verschae, J. Ruiz-del-Solar, M. Correa (2006). Gender Classification of Faces Using Adaboost, *11th Iberoamerican Congress in Pattern Recognition*, CIARP 2006 Cancun, Mexico, November 2006, 68–78.

[9] B. Jun, T. Kim and D. Kim (2011). A compact local binary pattern using maximization of mutual information for face analysis, *Pattern Recognition* Volume 44, Issue 3, 2011, 532–543.

[10] S.A. Khan, M. Ahmad, M. Nazir and N. Riaz (2014). A Comparative Analysis of Gender Classification Techniques, *Middle-East Journal of Scientific Research* 20(1):1–13.

[11] H.F. Alrashed and M.A. Berbar (2013). Facial Gender Recognition Using Eyes Images, *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 2, Issue 6, June 2013.

[12] A.M. Mirza, M. Hussain, H. Almuzaini, G. Muhammad, H. Aboalsamh and George Bebis (2013). Gender Recognition Using Fusion of Local and Global Facial Features. *In G. Bebis et al. (Eds.): ISVC 2013, Part II, LNCS 8034*, pp. 493–502, 2013.

[13] M. Rodrigues, M. Kormann and P. Tomek (2014). ROI Sensitivity Analysis for Real Time Gender Classification, submitted to *CSCC-2014 The 18th Int Conf of Circuits, Systems, Communications and Computers*, July 17–21, 2014, Greece.

[14] R. Zabih, J. Woodfill (1994). Non-parametric local transforms for computing visual correspondence. *In: ECCV. (1994)*, Secaucus, NJ, USA, Springer-Verlag New York, Inc. 151–158.

[15] R. Spangenberg, T. Langner, R. Rojas (2013). Weighted semi-global matching and center-symmetric census transform for robust driver assistance. *In Wilson, R., Hancock, E., Bors, A., Smith, W., eds.: Computer Analysis of Images and Patterns.* Volume 8048 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 34–41.

[16] B. Froba and A. Ernst (2004). Face detection with the modified census transform, *Proc. of IEEE 6th Int Conf on Automatic Face and Gesture Recognition*, 91–96.

[17] S.K. Gehrig, C. Rabe (2010). Real-Time Semi-Global Matching on the CPU. *In: CVPR Workshops*, San Francisco, CA, USA (June 2010) 85–92.

[18] T. Ojala, M. Pietikäinen, and D. Harwood (1994). Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, *Proceedings of the 12th IAPR International Conference on Pattern Recognition (ICPR 1994)*, vol. 1, 582–585.

[19] T. Ojala, M. Pietikäinen, and D. Harwood (1996). A Comparative Study of Texture Measures with Classification Based on Feature Distributions, *Pattern Recognition*, vol. 29, 51–59.

[20] T. Ahonen, A. Hadid, and M. Pietikäinen (2006). Face description with local binary patterns: Application to face recognition. *TPAMI*, 28(12):2037–2041, 2006.

[21] S. Belkasim (2011). Multi-resolution Analysis Using Symmetrized Odd and Even DCT Transforms, *Data Compression Conference (DCC)*, 447pp.

[22] K. Daewon and S. Daekyu (2003). Energy-based adaptive DCT/IDCT for video coding, *International Conference on Multimedia and Expo ICME'03*, Vol 1, 557–560.

[23] S. Gharge and S. Krishnan (2007). Simulation and Implementation of Discrete Cosine Transform for MPEG-4, *International Conference on Computational Intelligence and Multimedia Applications*, Vol 4, 137–141.

[24] Matlab Revision R2012b Documentation (2012). *Mathworks Online Documentation*, [Online] Available at www.mathworks.co.uk/help

[25] A. Webb and K. Copsey (2011). *Statistical Pattern Recognition*, 3rd edition, Wiley, 666pp.

[26] FEI Face Database (2014). [Online] Available http://fei.edu.br/ cet/facedatabase.html

A New Approach for Color Image Segmentation with Hierarchical Adaptive Kernel PCA

R. Kountchev, Noha A. Hikal, and R. Kountcheva

Abstract— A new approach for color image segmentation is presented, based on the algorithm for hierarchical adaptive Kernel Principal Component Analysis (HAKPCA). It permits to enhance the accuracy of the color segmentation in the cases when the vectors distribution in the color space is not Gaussian. This is achieved through applying the polynomial kernel used for the nonlinear transform of the RGB color space, after which on the expanded color vectors is applied the HAKPCA algorithm. In result is obtained high decorrelation of the transformed color vectors and concentration of the basic information in their first 2 components. This, on the other hand, permits to reduce the number of the transformed vectors components, retaining the first 2 only. In the new two-dimensional color space the color vectors are clustered and could be classified with high accuracy (by using k-nearest neighbors, SVM, LDA, , neural networks, etc), for example. One more advantage of the offered HAKPCA-based approach for color segmentation is the reduced computational complexity of the algorithm and the convenience for parallel implementation, which is of high importance for its real-time applications of any kind. The new approach could be used as a basis for medical image segmentation, melanoma recognition, and skin color segmentation.

***Keywords*— color segmentation, color space reduction, Kernel Principal Component Analysis, Hierarchical Adaptive Kernel Principal Component Analysis.** .

I. Introduction

THE color image segmentation is of high significance in computer vision as the first stage of the processing, concerning detection and extraction of objects with predefined color, shape of the visible part of the surface, and texture. The existing color image segmentation techniques can be classified into seven main approaches based on: edge detection, region growing, neural network based, fuzzy logic, histogram analysis, Support Vector Machine and principal color [1-5].

One of the contemporary methods for color image segmentation is the adaptive models in the perceptual color space, based on neural networks as multilayer perceptrons with multi-sigmoid activation function [6]. Recently special attention attracted the methods for human skin segmentation in color images [6-15]. These methods are mainly based on different color spaces, adaptive color space switching, skin color models and detection techniques.

The color space representation based on the KLT (PCA) [16,17,18] offers significant advantages in the efficient image processing, as image compression and filtration, color segmentation, etc. In this paper, a new approach for adaptive object color segmentation is presented through combining the linear and nonlinear Principal Component Analysis (PCA). The basic problem of PCA, which makes its application for efficient representation of the image color space relatively difficult; is related to the hypothesis for Gaussian distribution of the primary RGB vectors. One of the possible approaches for solving the problem is the use of PCA variations, such as: the nonlinear Kernel PCA (KPCA) [19,20], Higher-Order SVD (HOSVD) [21], Probabilistic PCA (PPCA) [22], Independent Component Analysis (ICA) [23], etc. In this work, for the color space representation is used an adaptive method for transform selection: linear PCA or nonlinear KPCA. The first transform (the linear PCA) could be considered as a particular case of the KPCA. The linear PCA is carried out on the basis of the already described adaptive color PCA (ACPCA)[15]. The choice of ACPCA or KPCA is made through evaluation of the kind of distribution of the vectors, which describe the object color: Gaussian or not.

The paper comprises the following parts: Section 2 - Description of the Color Kernel PCA, Section 3 - the algorithm for color image segmentation by using HAPCA, Section 4 - the algebraic calculation of eigen images through APCA with a 3×3 matrix, Section 5 – the evaluation of the color vectors distribution in the transformed space, Section 6 – experimental results, and Conclusions

R. K. Kountchev is with the Department of Radio Communications and Video Technologies at the Technical University of Sofia, Bulgaria (phone:+3592 979 0219 e-mail: rkountch@tu-sofia.bg).

Noha A. Hikal is with the Department of Information Technology at the faculty of Computers and Information Systems, Mansoura University, Mansoura, Egypt. On loan to the university of Taibah, faculty of sciences computer engineering, Madinah, KSA. (corresponding author phone:+2-0100-4062620; e-mail:dr.nahikal@mans.edu.eg, or nmhikal_5@yahoo.com).

R. A. Kountcheva is with T&K Engineering Co., Sofia, Bulgaria (e-mail: kountcheva_r@yahoo.com).

II. Color Kernel PCA

In the general case, with KPCA performs a nonlinear transform (extension) of the original centered vectors $\vec{X}_s$ over S pixels ($\vec{X}_s = \sum_{s=1}^{S} \vec{X}_s$) into a high-dimensional space, and then, for the obtained transformed vectors $\Phi(\vec{X}_s)$ the PCA is applied. The aim is in the new, multidimensional space, the vertices of vectors $\Phi(\vec{X}_s)$ to be concentrated in an area, which is accurately enough enveloped by a hyperellipsoid, whose axes are the eigenvectors of the covariance matrix of vectors $\Phi(\vec{X}_s)$. Fig.1 illustrates the idea of the new multidimensional color space[15].

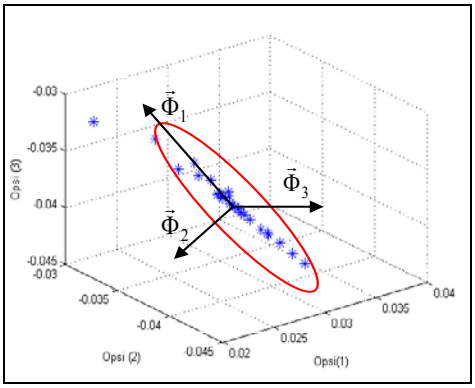


Fig. 1. Plot of skin color samples in $\vec{\Phi}_1, \vec{\Phi}_2, \vec{\Phi}_3$ space of the ACPCA

In particular, it is possible for the obtained vectors $\Phi(\vec{X}_s)$ in the transformed space to be represented by their projections on the first eigenvector $\vec{v}_1$ of their covariance matrix, as shown in Fig. 2. For the example, shown in this figure, the eigenvector $\vec{v}_1$ is projected the basic part from the multitude of all transformed vectors $\Phi(\vec{X}_s)$. The original 3D color vectors $\vec{C}_s$ are first centered:

$$\vec{X}_s = \vec{C}_s - \vec{m}_C \quad \text{for} \quad s = 1,2,..,S \tag{1}$$

Where $\vec{m}_C$ is the mean value of the color vector and then follows some kind of nonlinear transform, which uses the selected nonlinear function Φ(.). In result, the corresponding N-dimensional vectors, $\Phi(\vec{X}_s)$ (N≥3) are obtained. The value of N depends on the selected function Φ(.), used for the nonlinear transform [19,20].

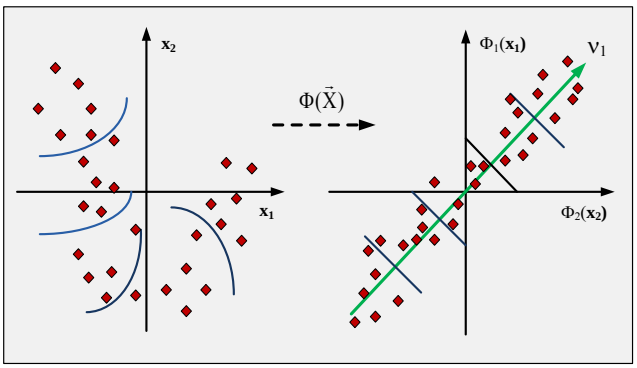


Fig. 2. Color space transform with KPCA

The covariance matrix $[\widetilde{K}_x]$ of the transformed color vectors $\Phi(\vec{X}_s)$ is of size N×N and can be computed mathematically as:

$$[\widetilde{K}_x] = \frac{1}{S}\sum_{s=1}^{S} \Phi(\vec{X}_s).\Phi(\vec{X}_s)^t = E\{\Phi(\vec{C}_s - \vec{m}_c).\Phi(\vec{C}_s - \vec{m}_c)^t\}, \tag{2}$$

Where: $\Phi(\vec{X}_s) = [\Phi(x_{s1}), \Phi(x_{s2}), .., \Phi(x_{sN})]^t$ for s = 1,2,..,S.

For each eigenvalue $\widetilde{\lambda}_i$ and eigenvector $\vec{v}_i = [v_{i1}, v_{i2}, .., v_{iN}]^t$ of the matrix $[\widetilde{K}_x]$ is performed the relation:

$$[\widetilde{K}_x]\vec{v}_i = \widetilde{\lambda}_i \vec{v}_i \qquad \text{for i = 1,2,..,N.} \tag{3}$$

After substituting in (3) using (2), results in:

$$[\widetilde{K}_x]\vec{v}_i = \frac{1}{S}\sum_{s=1}^{S} \Phi(\vec{X}_s)\Phi(\vec{X}_s)^t \vec{v}_i = \widetilde{\lambda}_i \vec{v}_i \tag{4}$$

In result of the transformation of (4), known as the "kernel trick" [19]. Therefore, the i^{th} eigenvector can be obtained as:

$$\vec{v}_i = \frac{1}{S\widetilde{\lambda}_i}\sum_{s=1}^{S} (\Phi(\vec{X}_s)^t.\vec{v}_i)\Phi(\vec{X}_s) = \sum_{s=1}^{S} \alpha_{si}\Phi(\vec{X}_s), \tag{5}$$

where for $\widetilde{\lambda}_i \neq 0$ the coefficient $\alpha_{si} = \frac{\Phi(\vec{X}_s)^t.\vec{v}_i}{S\widetilde{\lambda}_i}$.

From this follows, that:

$$[\widetilde{K}_x]\vec{v}_i = \widetilde{\lambda}_i \vec{v}_i = \widetilde{\lambda}_i \sum_{s=1}^{S} \alpha_{si}\Phi(\vec{X}_s). \tag{6}$$

Substituting (5) in (4), it is obtained:

$$(\frac{1}{S}\sum_{s=1}^{S} \Phi(\vec{X}_s)\Phi(\vec{X}_s)^t)(\sum_{l=1}^{S} \alpha_{il}\Phi(\vec{X}_l)) = \widetilde{\lambda}_i \sum_{l=1}^{S} \alpha_{li}\Phi(\vec{X}_l)$$

or

$$\frac{1}{S}\sum_{s=1}^{S}\sum_{l=1}^{S}\Phi(\vec{X}_s)\Phi(\vec{X}_s)^t\Phi(\vec{X}_l)\alpha_{il}=\tilde{\lambda}_i\sum_{l=1}^{S}\alpha_{li}\Phi(\vec{X}_l),$$

from which follows:

$$\sum_{s=1}^{S}\sum_{l=1}^{S}\Phi(\vec{X}_s)\Phi(\vec{X}_s)^t\Phi(\vec{X}_l)\alpha_{li}=S\tilde{\lambda}_i\sum_{l=1}^{S}\alpha_{li}\Phi(\vec{X}_l). \quad (7)$$

After multiplying the left side of the above equation with the vector $\Phi(\vec{X}_s)^t$, results in:

$$\sum_{s=1}^{S}\sum_{l=1}^{S}\Phi(\vec{X}_s)^t\Phi(\vec{X}_s)\Phi(\vec{X}_s)^t\Phi(\vec{X}_l)\alpha_{li} = S\tilde{\lambda}_i\sum_{l=1}^{S}\alpha_{li}\Phi(\vec{X}_l)^t\Phi(\vec{X}_s). \quad (8)$$

The dot product of vectors $\Phi(\vec{X}_s)$ and $\Phi(\vec{X}_l)$ can be represented through the kernel function $k(\vec{X}_s,\vec{X}_l)$, defined by the relation:

$$k(\vec{X}_s,\vec{X}_l)=\Phi(\vec{X}_s)^t.\Phi(\vec{X}_l) \quad \text{for s,}l\text{=1,2,..,S.} \quad (9)$$

Here, the term $k(\vec{X}_s,\vec{X}_l)$ represents the elements (s, *l*) of the Gram matrix [K] of size S×S, called "kernel matrix". After substituting (9) in (8), it could be represented as follows

$$[K]^2.\vec{\alpha}_i=S\tilde{\lambda}_i[K]\vec{\alpha}_i. \quad (10)$$

Under the condition, that the matrix [K] is positively defined (i.e. when it eigenvalues are positive) it could be presented shorter than (10). Then:

$$[K]\vec{\alpha}_i=S\tilde{\lambda}_i\vec{\alpha}_i. \quad (11)$$

From this relation follows, that $S\tilde{\lambda}_i$ are the eigenvalues of the matrix [K], and $\vec{\alpha}_i=[\alpha_{i1},\alpha_{i2},..,\alpha_{iS}]^t$ are the corresponding eigenvectors of same matrix. Taking into account the requirement $\vec{v}_i^t\vec{v}_i=1$ from (5) is obtained the relation:

$$\sum_{s=1}^{S}\sum_{l=1}^{S}\alpha_{li}\alpha_{si}\Phi(\vec{X}_l)^t.\Phi(\vec{X}_s)=1 \text{ or } \vec{\alpha}_i^t[K]\vec{\alpha}_i=1 \quad (12)$$

After substituting (11) in (12) is obtained $S\tilde{\lambda}_i\vec{\alpha}_i^t\vec{\alpha}_i=1$, from which is defined the square of the module of the vector $\vec{\alpha}_i=[\alpha_{i1},\alpha_{i2},..,\alpha_{iS}]^t$:

$$\left\|\vec{\alpha}_i\right\|^2=\vec{\alpha}_i^t.\vec{\alpha}_i=\sum_{s=1}^{S}\alpha_{si}^2=1/S\tilde{\lambda}_i. \quad (13)$$

In the general case, the vectors $\Phi(\vec{X}_s)$ in (9) are not centered. On order to apply the PCA on them, they should be centered in advance, and in result are obtained the vectors:

$$\breve{\Phi}(\vec{X}_s)=\Phi(\vec{X}_s)-E\{\Phi(\vec{X}_s)\}, \quad (14)$$

Where: $\vec{m}_{\Phi}=E\{\Phi(\vec{X}_s)=\frac{1}{S}\sum_{s=1}^{S}\Phi(\vec{X}_s)$.

The covariance matrix $[\breve{K}]$ of the centered vectors $\Phi(\vec{X}_s)$ is of size S×S and is defined by the relation:

$$[\breve{K}]=\frac{1}{S}\sum_{s=1}^{S}\breve{\Phi}(\vec{X}_s)^t.\breve{\Phi}(\vec{X}_l)=E(\breve{\Phi}(\vec{X}_s)^t.\breve{\Phi}(\vec{X}_l)). \quad (15)$$

The matrix kernel is:

$$\breve{k}(\vec{X}_s,\vec{X}_l)=\breve{\Phi}(\vec{X}_s)^t.\breve{\Phi}(\vec{X}_l)=(\Phi(\vec{X}_s)-\vec{m}_{\Phi})^t.(\Phi(\vec{X}_l)-\vec{m}_{\Phi})$$

$$\text{for s,}l\text{=1,2, . . ,S.} \quad (16)$$

The relation between covariance matrices $[\breve{K}]$ and [K] is:

$$[\breve{K}]=[K]-2[I_{1/s}][K]+[I_{1/s}][K][I_{1/s}], \quad (17)$$

where $[I_{1/s}]$ is a matrix of size S×S, whose elements are equal to 1/S.

The projection of the vector $\Phi(\vec{X}_s)$ on the eigenvector $\vec{v}_i$ in the S-dimensional space is:

$$\mathrm{Pr}_{si}=\Phi(\vec{X}_s)^t.\vec{v}_i=\sum_{s=1}^{S}\alpha_{is}\Phi(\vec{X}_i)^t.\Phi(\vec{X}_s)=\sum_{s=1}^{S}\alpha_{is}k(\vec{X}_i,\vec{X}_s)$$

$$\text{for i = 1,2,3,.,N.} \quad (18)$$

Using the projections Pr_{si} of the vector $\Phi(\vec{X}_s)$ on each of the first k ≤ N eigenvectors $\vec{v}_i$ (for i = 1,2,..,k) could be used in making the decision for the classification of the s[th] pixel to the dominant color of a selected object, using some of the well-known classifiers, as: SVM, LDA, k-nearest neighbors, neural networks, etc [24].

To carry out the KPCA could be used different kinds of kernel functions, such as the polynomial, the Gaussian, the sigmoid, etc. By substituting $\Phi(\vec{X}_s)=\vec{\mathrm{x}}$ and $\Phi(\vec{X}_l)=\vec{\mathrm{y}}$ the polynomial kernel function of degree d is defined by the relation:

$$\mathrm{k}(\vec{\mathrm{x}},\vec{\mathrm{y}})=(\vec{\mathrm{x}}^{\mathrm{t}}.\vec{\mathrm{y}})^{\mathrm{d}}. \quad (19)$$

For d=2 and if assumed that for the transformation of the 3D vectors $\vec{X}_s=[\mathrm{x}_{s1},\mathrm{x}_{s2},\mathrm{x}_{s3}]^{\mathrm{t}}$ and $\vec{X}_l=[\mathrm{x}_{l1},\mathrm{x}_{l2},\mathrm{x}_{l3}]^{\mathrm{t}}$ into N-dimensional is used the nonlinear function Φ(.), then:

$$\vec{\mathrm{x}}=\Phi(\vec{\mathrm{X}}_s)=[\Phi_{s1},\Phi_{s2},..,\Phi_{sN}]^{\mathrm{t}},$$

$$\vec{\mathrm{y}}=\Phi(\vec{\mathrm{X}}_1)=[\Phi_{l1},\Phi_{l2},..,\Phi_{lN}]^t \quad (20)$$

where the vectors components are defined by the relations:

$$\Phi_{si} = x_{s_i p_1}^{r_1} x_{s_i p_2}^{r_2}, \Phi_{li} = x_{l_i p_1}^{r_1} x_{l_i p_2}^{r_2} \quad (21)$$

for r1, r2 = 0,1, p1, p2 = 1,2,3, i = 1,2,..N and s,*l* =1,2,..,S.

In this case the maximum value of N is N = 9. In order to reduce the needed calculations, is suitable to use smaller number of the possible 9 components of the quadratic function Φ(.).

For example, if assumed N = 3 and if only mixed products of the vectors components $\vec{X}_s$ and $\vec{X}_l$ are chosen, from (21) follows:

$$\vec{x} = \Phi(\vec{X}_s) = [x_{s1}x_{s2}, x_{s1}x_{s3}, x_{s2}x_{s3}]^t,$$
$$\vec{y} = \Phi(\vec{X}_1) = [x_{l1}x_{l2}, x_{l1}x_{l3}, x_{l2}x_{l3}]^t. \quad (22)$$

Then the corresponding kernel function of vectors $\Phi(\vec{X}_s)$ and $\Phi(\vec{X}_l)$ is represented by the polynomial below:

$$\begin{aligned} k(\vec{x},\vec{y}) &= [\Phi_{s1}, \Phi_{s2}, \Phi_{s3}]^t . [\Phi_{l1}, \Phi_{l2}, \Phi_{l3}] \\ &= x_{s1}x_{s2}x_{l1}x_{l2} + x_{s1}x_{s3}x_{l1}x_{l3} + x_{s2}x_{s3}x_{l2}x_{l3}. \end{aligned} \quad (23)$$

In particular, for d=1, $\Phi(\vec{X}_s) = \vec{X}_s$ and $\Phi(\vec{X}_l) = \vec{X}_l$ the corresponding kernel function is linear:

$$\begin{aligned} k(\vec{x},\vec{y}) &= [x_{s1}, x_{s2}, x_{s3}]^t . [x_{l1}, x_{l2}, x_{l3}] \\ &= x_{s1}x_{l1} + x_{s2}x_{l2} + x_{s3}x_{l3}. \end{aligned} \quad (24)$$

From the above, it follows that KPCA is transformed into linear PCA (i.e. PCA is a particular case of KPCA).

III. Algorithm for Color Image Segmentation by Using HAPCA

The general algorithm for objects segmentation in the extended color space, based on the Kernel HAPCA (KHAPCA) and a classifier of the reduced vectors, is given in Fig. 3.

In the block for preprocessing, each color vector $\vec{C}_s = [R_s, G_s, B_s]^t$ is transformed into the corresponding expanded vector $\vec{P}_s$. If the chosen kernel-function is polynomial, and the 3-dimensional color space is transformed into a 9-dimensional, then the components p_{is} of the vectors $\vec{P}_s$ could be defined as follows:

$$\begin{aligned} \vec{P}_s &= [R_s, G_s, B_s, R_s^2, G_s^2, B_s^2, R_s G_s, B_s G_s, R_s B_s,]^t \\ &= [P_{1s}, P_{2s}, P_{3s}, P_{4s}, P_{5s}, P_{6s}, P_{7s}, P_{8s}, P_{9s},]^t \end{aligned}$$

for s=1,2,..,S.

In order to put all components p_{is} in the range [0, 255], these with a consecutive number i=4,5,..,9 products $R_s^2, G_s^2, B_s^2, R_s G_s, B_s G_s, R_s B_s$ are quantized in the range 0 -255. The vectors $\vec{P}_s$ are then transformed by the 2-level HAPCA, whose algorithm is shown in Fig. 3.

As a result of the transform are obtained the 2-component vectors $\vec{E}_s = [E_{1s}, E_{2s}]^t$, which are used to substitute the input 9-components vectors

$$\vec{P}_s = [P_{1s}, P_{2s}, P_{3s}, P_{4s}, P_{5s}, P_{6s}, P_{7s}, P_{8s}, P_{9s},]^t .$$

In this way the performance of the classifier is also simplified, because it have to process the vectors $\vec{E}_s$ in the 2-dimensional, instead of the 9-dimensional space. At its output are separated (indexed) all pixels in the image, whose corresponding vectors $\vec{E}_s$ are in the area of the cluster, belonging to the object. With this the color segmentation is finished.

In accordance with the algorithm shown in Fig. 4, for the 2-level HAPCA [25], the 9 components of each input vector $\vec{P}_s$ are divided into 3 groups, which contain the 3-components vectors $\vec{P}_{1s} = [P_{11s}, P_{12s}, P_{13s}]^t$, $\vec{P}_{2s} = [P_{21s}, P_{22s}, P_{23s}]^t$ and $\vec{P}_{3s} = [P_{31s}, P_{32s}, P_{33s}]^t$.

In the HAPCA first level, each group of 3-dimensional vectors $\vec{P}_{ks} = [P_{k1s}, P_{k2s}, P_{k3s}]^t$ for k=1,2,3 is performed Adaptive PCA (APCA) with a transform matrix of size 3×3. The so obtained vectors from each group comprise 3 "eigen" images, shown in Fig. 3 with different colors. These images are rearranged in accordance to the rule:

$$Pow_1 \geq Pow_2 \geq,..., Pow_9, \quad (25)$$

where

$$Pow_l = \frac{1}{QR} \sum_{i=1}^{Q} \sum_{j=1}^{R} p_{i,j}^2(l) \text{ for l=1,2,..,9}$$

is the power of each component l of the nonlinear transformed color image of size Q×R=S and pixels $p_{i,j}(l)$. After that these components are separated again, this time into 3 groups, of 3 images each. The vectors, obtained from pixels with same coordinates in the images from each group, are of 3 components.

For the second level, of HAPCA for each group of 3-dimensional vectors is performed ACPCA with a transform matrix of size 3×3 [15]. The so obtained vectors from each group build the 3 eigen images. These images are rearranged again in accordance with (25). As a result, are obtained the 9 eigen images E_1-E_9, from which are retained the first two (E_1 and E_2) only, which carry the main information, needed for the color objects segmentation. As a result, the computational complexity of HAPCA is smaller than that of PCA, for the case, when it is used to transform directly the 9-dimensional vectors $\vec{P}_s$. In this way, the general computational complexity of HAPCA and a classifier, needed for the processing of the vectors $\vec{P}_s$ is lower than that, needed for the processing of same vectors with PCA and a classifier. From the pixels with same coordinates in the images E_1 and E_2 are obtained the vectors $\vec{E}_s = [E_{1s}, E_{2s}]^t$, which are then used by a classifier.

ISBN: 978-1-61804-236-1

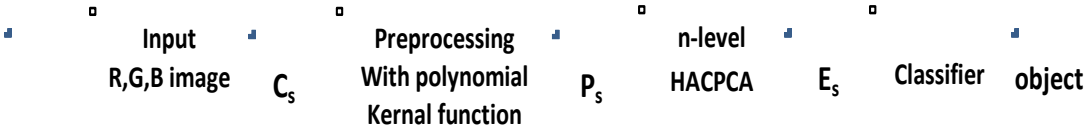


Fig. 3. Block diagram of the algorithm for image segmentation in the expanded color space

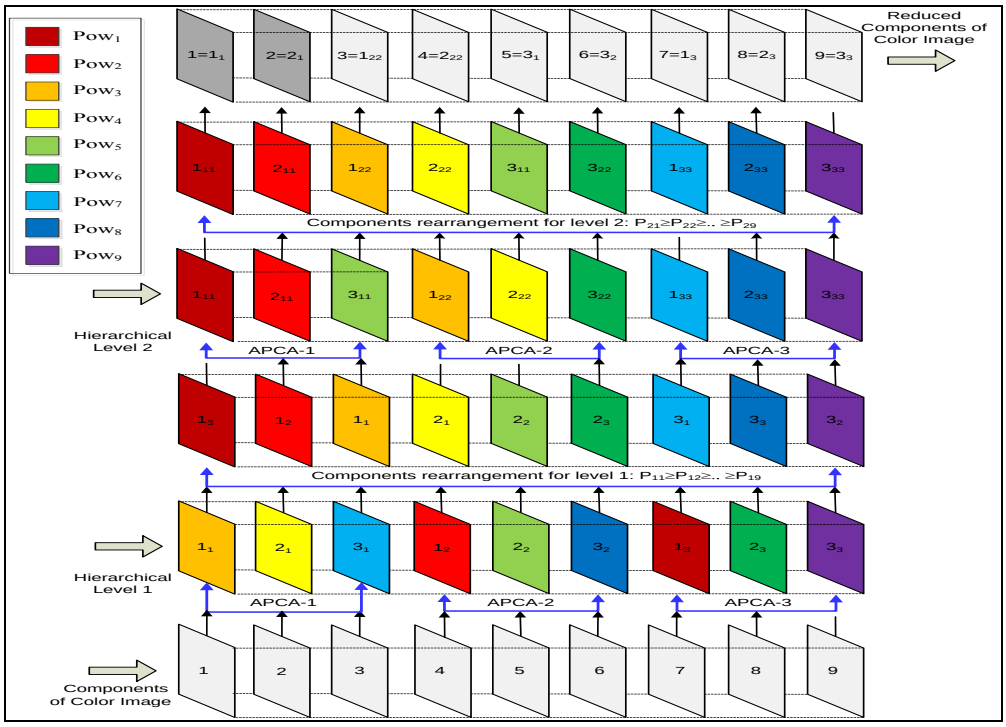
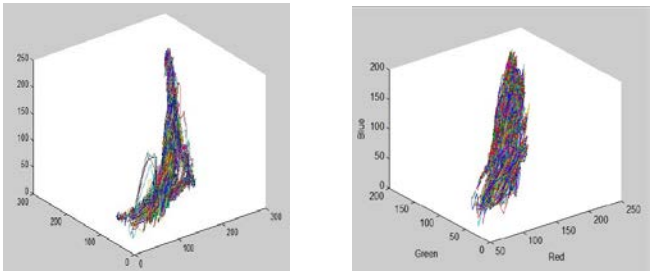


Fig. 4. Algorithm for 2 levels Hierarchical APCA for color image

VI. EXPERIMENTAL RESULTS

To verify the feasibility of the proposed algorithm, skin pigmentation images were tested and evaluated. Fig. 5, 6 show the original tested images and their color vectors distribution in RGB space, respectively. It can be seen clearly that their color distributions are considered a non-linear Gaussian ones. These images are passed through the HAPCA algorithm (presented in Fig. 3 and 4). The obtained transformed vectors $\vec{E}_s$ in the new color space E_{1s},E_{2s},E_{3s} are plotted in 3D domain shown in Fig.7. It can be noticed that the proposed techniques is able to concentrate the energy of the different skin color into very small and close components of transformed vectors.

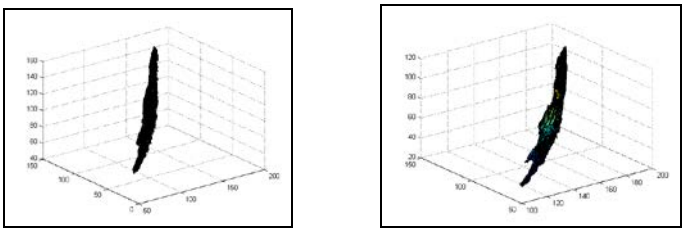
Fig. 5 Original skin pigmentation images



(a) (b)

Fig.6 (a) and (b) are the corresponding RGB distribution for original images Fig.5-a, and 5-b respectively

(a) (b)



(a) (b)

ISBN: 978-1-61804-236-1

Fig. 7 Plots of HAPCA transformed Vectors for Fig.5-a,b respectively

The HPCA transformed coefficients are then used to train a classifier. For briefing, fuzzy K-means clustering is used. The segmentation results are shown in Fig. 8 a, b respectively.

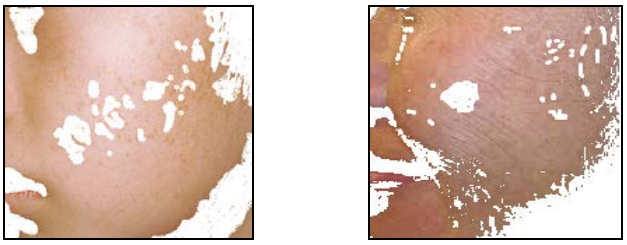


Fig. 8 Skin color segmentation based on HAPCA

The proposed approach depends mainly on the evaluation of the kind of the color vectors distribution. On the basis of color distribution, the more efficient transform. For a non-Gaussian distribution of the vectors, the KPCA is used. The selected nonlinear transform results in negligible expansion of the original color space, which increases slightly the number of needed calculations.

The main advantage of the new approach for color space representation is that in result of its adaptation in respect to color vectors distribution, it could be used as universal tool for efficient image processing. One more advantage of HAPCA towards the PCA is the low computational complexity.

On the basis of the presented approach was developed the new algorithm for objects color segmentation, distinguished by its high accuracy. This algorithm could be used in the CBIR systems for extraction of objects with preset color, in the computer vision systems for detection and tracking of objects in correspondence to their color under changing surveillance conditions, for automatic control of various manufacturing processes, etc.

REFERENCES

[1] X. Jie, S. Fei, "Natural color image segmentation", Proc. of IEEE IC on Image Processing (ICIP'03), Barcelona, Spain 2003, pp. 973-976.

[2] E. Navon, O. Miller, A. Averabuch, " Color image segmentation based on adaptive local thresholds", Image and Vision Computing, 23, 2005, pp. 69-85.

[3] K. Deshmukh, G. Shinde, "An adaptive color image segmentation, Electronic Letters on Computer Vision and Image Analysis, 5 (4), 2005, pp. 12-23.

[4] Z. Yu, H.-S. Wong, G. Wen, "A modified support vector machine and its Application to image segmentation", Image and Vision Computing, 29, 2011, pp. 29-40.

[5] X. Wang. T. Wang, J. Bu," Color image segmentation using pixel wise support vector machine classification", Pattern Recognition, 44, 2011, pp. 777-787.

[6] K. Bhoyar, O. Kakde, "Skin color detection model using neural networks and its Performance evaluation". Journal of Computer Science, 6 (9), 2010, pp. 963-968

[7] S. Phung, A. Bouzerdoum, D. Chai, "Skin Segmentation Using Color Pixel Classication: Analysis and Comparison", IEEE Transactions on Pattern Analysis and Machine Intelligence, January 2005, Vol. 27, No 1, pp. 148-154.

[8] V. Vezhnevets, V. Sazonov, A. Andreeva, "A Survey on Pixel-based Skin Color Detection Techniques", GRAPHICON'03, 2003, pp. 85–92.

[9] H. Stern, B. Efros, "Adaptive Color Space Switching for Face Tracking in Multi-colored Lighting Environments". Proc. of the Intern. Conference on Automatic Face and Gesture Recognition, 2002, pp. 249-255.

[10] F. Tomaz, T. Candeias, H. Shahbazkia, "Improved Automatic Skin Detection in Color Images". Proc. VIIth Digital Image Computing: Techniques and Applications, C. Sun, H. Talbot, S. Ourselin, T. Adriaansen (Eds.), Sydney, Dec. 2003, pp. 10-12.

[11] P. Kakumanu, S. Makrogiannis, N. Bourbakis, "A Survey of Skin-color Modeling and Detection Methods", Pattern Recognition, Elsevier, 40, 2007, pp. 1106-1122.

[12] M. Ionita, P. Corcoran, "Benefits of Using Decorrelated Color Information for Face Segmentation/tracking". Advances in Optical Technologies, Hindawi Publishing Corporation, 2008, ID 583687.

[13] J. Lee, S. Yoo, "An Elliptical Boundary Model for Skin Color Detection". Proc. of the Intern. Conference on Imaging Science, Systems and Technology (CISST'02), 2002.

[14] R. Hassanpour, A. Shahbahrami, St. Wong, "Adaptive Gaussian Mixture Model for Skin Color Segmentation". World Academy of Science, Engineering and Technology, 41, 2008, pp. 1-6.

[15] Noha A. Hikal, R. Kountchev, Skin color segmentation using adaptive PCA and modified elliptic boundary model. International Proc. of the IEEE International Conference on Advanced Computer Science and Information Systems (IEEE ICACSIS'11), Jakarta, Universitas Indonesia, December 2011, pp. 407 - 412.

[16] A. Abadpour and S. Kasaei, "Principal Color and Its Application to Color Image Segmentation". Scientia Iranica, Vol. 15, No. 2, April 2008, pp 238-245.

[17] R. Dony, "Karhunen-Loève transform", Book Chapter in: The Transform and Data Compression Handbook, K. Rao, P. Yip, (Eds.), CRC Press LLC, 2001.

[18] R. Kountchev, R. Kountcheva, "Image color space transform with enhanced KLT". Book chapter in: New Advances in Intelligent Decision Technologies, K. Nakamatsu, G. Wren, L. Jain, R. Howlett (Eds.), Springer-Verlag, 2009, pp. 171-182.

[19] B. Scholkopf, A. Smola, K. Muller, "Kernel Principal Component Analysis. In: Advances in Kernel Methods–Support Vector Learning", B. Scholkopf, C. Burges, A. Smola (Eds.), MIT Press, Cambridge, MA, 1999, pp. 327-352.

[20] B. Scholkopf, A. Smola, and K. Muller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation", 10, 1998, pp. 1299-1319.

[21] A. Rovid, L. Szeidl, P. Varlaki, "The HOSVD Based Domain and the Related Image" Processing Techniques, Intern. Journal of Applied Mathematics and Informatics, Issue 3, Vol. 5, 2011, pp. 157-164.

[22] M. Tipping, C. Bishop," Probabilistic Principal Component Analysis", Journal of the Royal Statistical Society, Series B, 61, Part 3, 1999, pp. 611–622.

[23] A. Hyvärinen, E. Oja," Independent Component Analysis: Algorithms and Applications". Neural Networks, 13 (4-5), 2000, pp. 411-4301.

[24] J. Lee, M. Verleysen, "Nonlinear Dimensionality Reduction", Springer, 2007.

[25] R. Kountchev, R. Kountcheva, "One approach for decorrelation of multispectral images, based on Hierarchical Adaptive PCA". 12th WSEAS Intern. Conf. on Signal Processing, Computational Geometry and Artificial Vision (ISCGAV'12), Istanbul, Turkey, August 2012, pp. 68-73.

Compressive Sensing-based Target Tracking for Wireless Visual Sensor Networks

Salema Fayed[a], Sherin Youssef[a], Amr El-Helw[b], Mohammad Patwary[c], and Mansour Moniri[c]
[a] Computer Engineering Department, [b] Electronics and Communication Department
College of Engineering and Technology
AAST, Alexandria, Egypt
[c] Faculty of Computing, Engineering and Technology
Staffordshire University, Stoke on Trent, UK

***Abstract*—**Limited storage, channel bandwidth, and battery lifetime are the main concerns when dealing with Wireless Visual Sensor Networks (WVSNs). Surveillance application for WVSNs is one of the important applications that requires high detection reliability and robust tracking, while minimizing the usage of energy as visual sensor nodes can be left for months without any human interaction. In surveillance applications, within WVSN, only single view target tracking is achieved to keep minimum number of visual sensor nodes in a 'wake-up' state to optimize the use of nodes and save battery life time, which is limited in WVSNs. Least Mean square (LMS) adaptive filter is used for tracking to estimate target's next location. Moreover, WVSNs retrieve large data sets such as video, and still images from the environment requiring high storage and high bandwidth for transmission which are limited. Hence, suitable representation of data is needed to achieve energy efficient wireless transmission and minimum storage. In this paper, the impact of CS is investigated in designing target detection and tracking techniques for WVSNs-based surveillance applications, without compromising the energy constraint which is one of the main characteristics of WVSNs. Results have shown that with compressive sensing (CS) up to 31% measurements of data are required to be transmitted, while preserving the detection and tracking accuracy which is measured through comparing targets trajectory tracking.

***Keywords*—** Compressive sensing, LMS, Surveillance applications, Target tracking, WVSN

I. Introduction

Wireless Visual Sensor Networks (WVSNs) have gained significant importance in the last few years and have emerged in several distinctive applications [1], [2]. Due to the evolvement of new technologies and techniques, there are immediate needs for automated energy-efficient surveillance systems. WVSN has targeted various surveillance applications in commercial, law enforcement and military purpose as well as traffic control, security in shopping malls and amusement parks. Systems have been developed for video surveillance including highway, subway and tunnel monitoring, in addition to remote surveillance of human activities such as elderly or patients care.

Visual sensor nodes are resource constraint devices bringing the special characteristics of WVSNs such as energy, storage and bandwidth constraints which introduced new challenges [3]. In WVSN large data sets such as video, and still images are to be retrieved from the environment requiring high storage and high bandwidth for transmission. Higher complexity of data processing and analysis is also challenging which are all quite costly in terms of energy consumption. Furthermore, wireless channels in surveillance applications are subject to noisy conditions; therefore, detection and tracking reliability within such resource constrained condition is the main challenge when designing WVSN surveillance applications. Energy efficient processing and efficient compression techniques are the strongest candidates to overcome such constrains while transmitting data for WVSN applications and hence minimize energy expenditure [2], [4].

Much work is present in the literature for surveillance applications within WVSNs [5], [6], [7]. Moreover, there is significant literature for target tracking surveillance applications in WVSN. Kalman filtering [8], [9] is relatively the best linear estimator for target tracking. Kalman filters are robust under optimal conditions, otherwise adaptive approaches are needed to solve these problems which can be either computationally expensive or not always be applicable in real time tracking.

Classical active contour [10] for target tracking fails in tracking multiple targets at once so occlusion problems are difficult to solve. In [11], the active contour is modified to resolve occlusion problem by performing merging and splitting when two targets get close together or move apart. However, there is a probability that the target is lost if the displacement of the target between two consecutive frames is large. Least Mean Sqaure (LMS) algorithm is relatively simple, has much lower computational complexity than the original Kalman filters and other adaptive algorithms; it does not require correlation function calculation nor does it require matrix inversions. Moreover, it is suitable for real time image applications [12], [13].

Based on the above literature, to attain a trade off between computational complexity and detection and tracking accuracy in the context of energy constrained WVSN, an image processing scheme is required with optimal pre-processing and post-processing can provide intended target detection and tracking accuracy within energy constraint nature of WVSN. Moreover, high volume data sets acquired in WVSN surveillance applications, should be represented in such a way that it requires optimum storage, energy, and allow reliable transmission due to the constraint on the physical and radio resources. In a surveillance application within WVSN, an image is captured and required to be sampled for storage as well as to be transmitted through wireless channel. According to Shannon-Nyquist sampling theory the minimum number of samples required to accurately reconstruct the signal without losses is twice its maximum frequency [14]. It is always challenging to reduce this sampling rate as much as possible, hence reducing the computation energy and storage. Recently proposed Compressive Sensing (CS) [14] is expected to be a strong candidate to overcome the above mentioned limitations where CS has been considered for different aspects of surveillance applications due to its energy efficient and low power processing as reported in [15], [16]..

CS theory shows that a signal can be reconstructed from far fewer samples than required by Nyquist theory as it is always challenging to reduce the sampling rate as possible, provided that the signal is sparse (where most of the signal's energy is concentrated in few non-zero coefficients) or compressible in some basis domain [17].

In [15], compressive sensing for background subtraction and multi-view ground plane target tracking are proposed. A convex optimization known as basis pursuit or orthogonal matching pursuit is exploited to recover only the target in the difference image using the compressive measurements to eliminate the requirement of any

ISBN: 978-1-61804-236-1

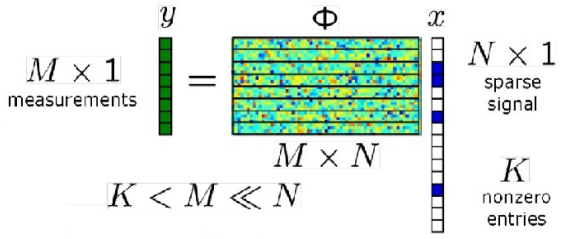


Fig. 1. Compressive sensing measurement process

auxiliary image reconstruction. Other work in compressive sensing for surveillance applications has been proposed in [18], where an image is projected on a set of random sensing basis yielding some measurements. In this paper, the impact of CS is investigated in designing target detection and tracking techniques for WVSNs-based surveillance applications, without compromising the energy constraint which is one of the main characteristics of WVSNs. CS is expected to reduce the size of sampled data with low complexity processing due to its low power simple process [17], hence saving space, energy of processing and transmission as well as channel bandwidth. Hence, a compressive sensing-based single/multi target tracking using LMS is proposed which is expected to reduce energy consumption, space requirement and communication overhead, with acceptable tracking reliability which will be represented as minimal mean square error (MSE).

The rest of the paper is organized as follows, Introduction to CS is presented in Section II. Section III presents the proposed system model. The proposed technique for CS-based target tracking is given in Section IV. Simulations and results are provided in Section V and finally the conclusion in Section VI.

II. Compressive Sensing theory

Suppose image X of size $(N \times N)$ is K-sparse that either sparse by nature or sparse in $\mathbf{\Psi}$ domain, CS exploits the sparsity nature of frames, so it compresses the image using far fewer measurements [19], [17], [20]. Although, it is not necessary for the signal itself to be sparse but compressible or sparse in some known transform domain $\mathbf{\Psi}$ according to the nature of the image, smooth signals are sparse in the Fourier basis, and piecewise smooth signals are sparse in a wavelet basis. $\mathbf{\Psi}$ is the basis invertible Orthonormal function of size $(N \times N)$ driven from a transform such as the DCT, fourier, or wavelet, where $K \ll N$, that is, only K coefficients of $\boldsymbol{x}$ are nonzero and the remaining are zero, thus the K-sparse image $\boldsymbol{X}$ is compressible. CS then guarantees acceptable reconstruction and recovery of the image from lower measurements compared to those required by shannon-Nyquist theory as long as the number of measurements satisfies a lower bound depending on how sparse the image is. Hence, $\boldsymbol{X}$ can be recovered from measurements of size M where $M \geq K \log N \ll N$. Eq.(1) shows the mathematical representation of X

$$\mathbf{X} = \mathbf{\Psi S} \tag{1}$$

$\boldsymbol{S}$ contains the sparse coefficients of $\boldsymbol{X}$ of size $(N \times N)$, $s_i =< \boldsymbol{X}, \psi_i^T >= \psi^{TX}$, $S = \mathbf{\Psi}^{TX}$. The image is represented with fewer samples from $\boldsymbol{X}$ instead of all pixels by computing the inner product between $\boldsymbol{X}$ and $\mathbf{\Phi}$, namely through incoherent measurements $\boldsymbol{Y}$ in Eq.(2), where $\mathbf{\Phi}$ is a random measurement matrix of size $(M \times N)$ where $K << M << N$. Fig.1 shows the CS measurement process [21].
$y_{\mathbf{1}} =< x, \phi_{\mathbf{1}} >, y_{\mathbf{2}} =< \mathbf{x}, \phi_{\mathbf{2}} >, \cdots, \mathbf{y_m} =< x, \phi_{\mathbf{m}} >$.

$$Y = \mathbf{\Phi} X = \mathbf{\Phi \Psi} S = \mathbf{\Theta} S \tag{2}$$

Since $M < N$, recovery of the image X from the measurements Y is undetermined, However, if S is K-sparse, and $M \geq K \log N$ it has been shown in [17] that X can be reconstructed by ℓ_1 norm minimization with high probability through the use of special convex optimization techniques without having any knowledge about the number of nonzero coefficients of X, their locations, neither their amplitudes which are assumed to be completely unknown a priori [20], [19], [22]

$$\mathbf{min} \|\hat{X}\|_{\ell_1} \ \mathbf{subject\ to}\ \mathbf{\Phi}\hat{X} = Y \tag{3}$$

Convex optimization problem can be reduced to linear programming known as Orthogonal Matching Pursuit (OMP) which was proposed in [23] to handle the signal recovery problem. It is an attractive alternative to Basis Persuit (BP) [24] for signal recovery problems.The major advantages of this algorithm are its speed and its ease of implementation. As seen, the CS is a very simple process as it enables simple computations at the encoder side (sensor nodes) and all the complex computations for recovery of frames are left at the decoder side or BS.

III. System model

This work proposes a compressive sensing model which is expected to reduce space requirements and communication overhead with low processing complexity while preserving detection and tracking accuracy.

Consider for a surveillance application a WVSN model composed of V visual sensor nodes and one or more BS. Each sensor node i is required to capture images from a video sequence and detect the presence of objects. At the time where a sensor node enters a 'wake-up' state, the time reference for the frame count is assumed to be $t = 0$. Hence, a single snapshot at $t = 0$ is expected to be stored within the memory allocated at the sensor node; that is assumed to be the background for the intended target tracking; denoted as $X_{\mathbf{b}}$. The following frames are the subsequent captured frames $X_{\mathbf{t}}$ with $t > 0$. Hence, $X_{\mathbf{b}}$ and $X_{\mathbf{t}}$ are the background and test images respectively of size $(N \times N)$ each. Let us assume most features of the targets are known to the monitoring center. However, the existence and the location of targets are required for monitoring. The receiver or BS also has prior explicit information of the background. To achieve higher compression rates, the foreground target is extracted first by background subtraction resulting in the difference frame. Hence, assuring sparsity as the difference frame is always sparse regardless the sparsity nature of real frames. Within the image frame, The extraction of foreground target $X_{\mathbf{d}}$ is achieved at each sensor node where CS is then applied for transmission through the wireless channel. At the BS side, the receiver decompresses the received compressed data obtaining $\hat{X}_{\mathbf{t}}$ to predicts the intended target's next location for tracking. The system model for the proposed WVSN is shown in Fig. 2

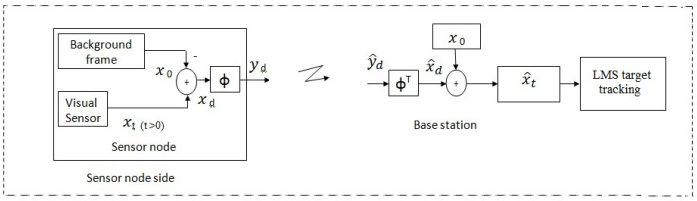


Fig. 2. The proposed model for WVSN-based surveillance application

IV. Proposed CS-based tracking algorithm

A. Compressive Sensing

At each sensor node, after each image frame is being captured, some preprocessing might be required. In our case, to assure sparsity

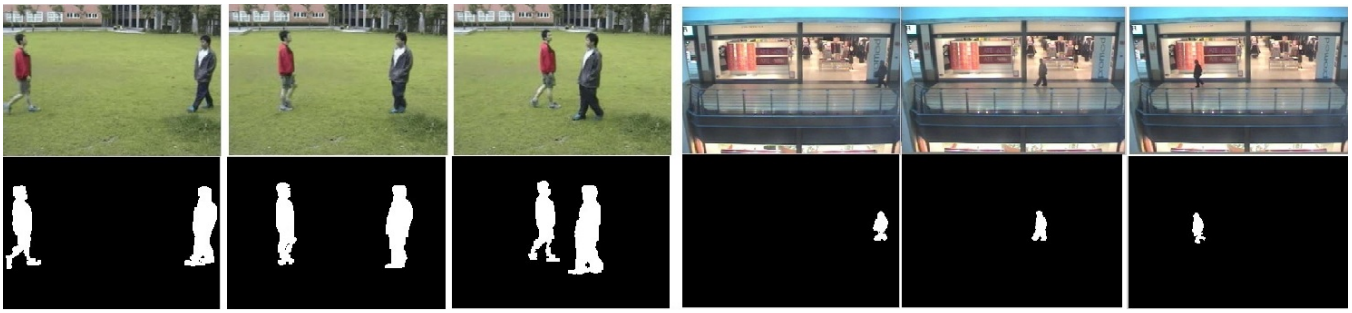


Fig. 3. First row shows test frames and background subtraction results in second row, the background subtracted frames are then compressed and used as a references to test detected location using CS. Left set of frames for scheme"1" (outdoor scheme) Right set of frames for scheme"2" (indoor scheme)

within the image frame, the foreground target is extracted first by background subtraction by subtracting $X_\mathbf{t}$ from $X_\mathbf{b}$ resulting in the difference frame $X_\mathbf{d}$. Hence, instead of producing the compressed measurements for $X_\mathbf{b}$ and $X_\mathbf{t}$ separately, the compressed measurements are produced directly for $X_\mathbf{d}$, as the difference frame is always sparse regardless of the sparsity nature of real frames. CS process is then applied to $X_\mathbf{d}$ by multiplying it by a random projection sensing matrix $\mathbf{\Phi}$ producing the compressed measurements $Y_\mathbf{d}$. At the BS side, the received compressed data is decompressed for the reconstruction of the estimated data $\hat{X}_\mathbf{d}$. As mentioned, $X_\mathbf{b}$ is known to the BS, making it possible to reconstruct the original test frame $\hat{X}_\mathbf{t}$ by adding $X_\mathbf{b}$ to $\hat{X}_\mathbf{d}$. Below are the steps undertaken during the entire process

- **Step 1:** $X_\mathbf{d} = |X_\mathbf{t} - X_\mathbf{b}| > Th$, where *th* is a given threshold to extract the foreground target
- **Step 2:** $\mathbf{\Phi}$ is a randomly chosen sensing matrix of size $M \times N$, where $M \ll N$
- **Step 3:** produce the compressed measurements $Y_\mathbf{d} = \mathbf{\Phi} X_\mathbf{d}$
- **Step 4:** sensor nodes transmits $Y_\mathbf{d}$ through the wireless channel
- **Step 5:** at the receiver side, $\mathbf{\Phi}$ must be known for the decompression of $Y_\mathbf{d}$. $\hat{X}_\mathbf{d}$ is reconstructed from the compressed measurements $Y_\mathbf{d}$, resulting in a frame with only the foreground target present.
- **Step 6:** the real frame $\hat{X}_\mathbf{t}$ is then obtained by adding $\hat{X}_\mathbf{d}$ to the background frame $X_\mathbf{b}$ which is also has to be known to the receiver side apriori.
- **Step 7:** the targets locations are obtained after reconstructing the real frame producing a trajectory for the complete path of each moving target

B. Least Mean Square (LMS) tracking

The LMS algorithm, is referred to as adaptive filtering algorithm since the statistics are estimated continuously, hence it can adapt to changes. LMS incorporates an iterative procedure during the training phase where it estimates the required coefficients to minimize the mean square error (MSE). This is accomplished through successive corrections to the expected set of coefficients which eventually leads to the minimum MSE. The LMS implementation process has been illustrated in Fig.(4).

As shown in Fig.4 the outputs are linearly combined after being scaled using corresponding weights. The weights are computed using LMS algorithm based on MSE criterion. Therefore the spatial filtering problem involves estimation of a signal from the received signal, by minimizing the error between the reference signal, which closely matches or has some extent of correlation with the desired signal estimate and the output. The LMS algorithm is initiated with an arbitrary value $\mathbf{w}(\mathbf{0})$ for the weight vector at $n = 0$. The successive corrections of the weight vector eventually leads to the minimum value of the mean squared error. The weight update can be given by

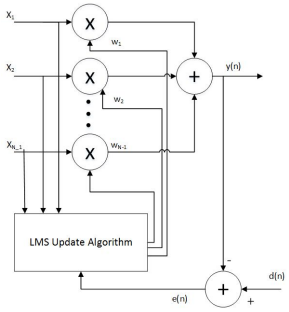


Fig. 4. An N-tap LMS adaptive filter

the following equation

$$w(n+\mathbf{1}) = w(n) + \mu \boldsymbol{x}(n)e(n) \tag{4}$$

where, $x(n)$ is the input signal, μ is the step size parameter, $e(n)$ is the MSE between the predicted output $y(n)$ and the reference signal $d(n)$ which is given by

$$e(n) = (d(n) - y(n))^{\mathbf{2}} \tag{5}$$

the output $\boldsymbol{y}(n)$ is calculated as follows

$$y(n) = x(n)w(n) \tag{6}$$

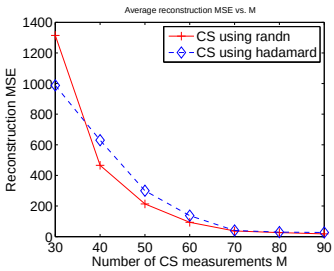
μ is selected by the autocorrelation matrix of the filter inputs.

V. Simulations and Results

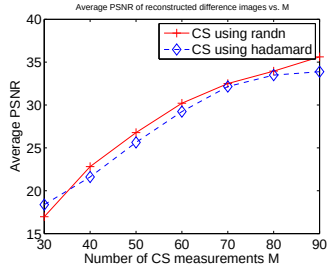
Based on the system model proposed, simulations and experiments are conducted to evaluate the performance of the CS-based target detection and tracking algorithm. Simulations are performed for the WVSN-based surveillance application in both outdoor and indoor scenes for single and multi-target tracking. Background and target's appearance are assumed to be static to investigate the effect of CS on the detection and tracking algorithms, hence schemes are chosen to reflect this assumption. Moreover, to illustrate the relation between the number of measurements required for CS to guarantee reconstruction and how sparse the image is. Simulations are performed on 2 different schemes with different sparsity levels; "outdoor scheme" is chosen to resemble an outdoor scenes for multi target tracking captured by [25]. While "indoor scheme" filmed for the EC funded CAVIAR project found in [26] for indoor scenes tracking a single target.

Mean square error (MSE) and peak signal to noise ratio (PSNR) are used as performance indicators to test the reliability of CS. MSE and PSNR are compared for different number of CS measurements *M*,

where the MSE is the reconstruction error measured between real and reconstructed frames and PSNR is measured after frames recovery to reflect the quality of image reconstruction which will later on reflects the ability of reliable tracking. The background frame and $\mathbf{\Phi}$ are known to the receiver node. Two candidate sensing matrices have been compared; normally distributed random numbers using Matlab function "randn" and a walsh-hadamard. Although the measurements are defined by a matrix multiplication, the operation of matrix-by-vector multiplication is seldom used in practice, because it has a complexity of O(MN) which may be too expensive for real time applications. When a randomly permutated Walsh-Hadamard matrix is used as the sensing matrix, the measurements may be computed by using a fast transform which has complexity of O(K log(N)) [27]. The Hadamard matrix, is an $(N \times N)$ square matrix whose entries are either +1 or -1 and whose rows are mutually orthogonal, the matrix is first randomly reordered then, M samples are randomly chosen to construct the $(M \times N)$ random sensing matrix $\mathbf{\Phi}$.



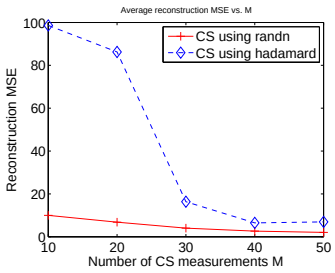
(a) Reconstruction MSE



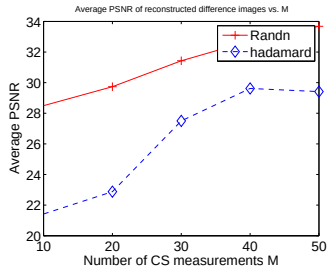
(b) PSNR

Fig. 5. Comparing reconstruction MSE and PSNR using randn and walsh sensing matrices for "outdoor scheme"

As stated earlier, the ability of reliable tracking depends on acceptable recovery of images. In other words, if CS fails in image reconstruction the targets location can not be detected. Hence, choosing the right value of M is critical in image reconstruction and afterwards tracking. It is clear from the results in Fig.5 and 6 for outdoor and indoor schemes respectively that for different sparsity levels different values of M and compression rates are required. When reaching optimum value of M least MSE while preserving a $33dB$ PSNR. For illustration, MSE decreases as M increases till reaching the optimum value, it has been shown that the lower bound on M is depending on how sparse the difference frame $X_{\mathbf{d}}$ is or in other words proportional to the ratio between the number of non-zero coefficients and the total number of pixels in a frame. For "outdoor scheme", CS sets M to 90 in Fig.5(a) to achieve satisfactory results. While for "indoor scheme", it is obvious from Fig.6(a) that for single-target tracking (where there is lower number of non-zero coefficients), better MSE is achieved with lower M, reduced to 50 for "indoor scheme"



(a) Reconstruction MSE



(b) PSNR

Fig. 6. Comparing reconstruction MSE and PSNR using randn and walsh sensing matrices for "indoor scheme"

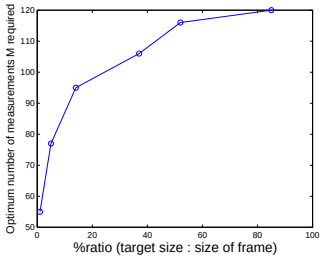
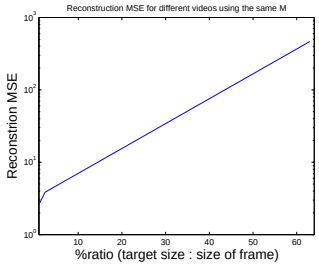


Fig. 7. Relation between the percentage ratio of target size:frame size vs. M

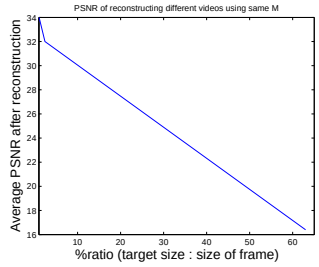
compared to multi-target tracking while maintaining least MSE and $33dB$ PSNR as in Fig.6.

As for MSE, Fig.5(b), 6(b) show the effect of M on PSNR for the different schemes. For each scheme, according to the sparsity nature of each scheme, the number of measurements M required will differ to obtain guaranteed reconstruction which is defined here in terms of PSNR. For low values of M it is hard to achieve a good PSNR, to reach the acceptable value, M should increase till reaching its optimum value as discussed earlier. To illustrate this for the "indoor scheme", to achieve a PSNR of $\approx 33dB$ M reached 50, while for the "outdoor scheme" if the same M is used, we could not attain a PSNR higher than $25dB$.

The above simulation were carried out using two different sensing matrices, Randn and walsh-Hadamard. They are compared with respect to MSE and PSNR as in Fig.5 and 6. It is clear from the results that when reaching the optimum value of M both sensing matrices perform nearly the same except in some cases in Fig.6 shows that Randn gives slightly a better performance than Hadamard. But this can be negligible when compared to the reduction in complexity gained by using Hadamard matrix which helps in accomplishing the



(a)



(b)

Fig. 8. Relation between the percentage ratio of target size:frame size and (a) reconstruction MSE, (b) average PSNR

main objective to save sensor nodes power and as a result maximizes their lifetime.

Fig.7 and 8 summarize and demonstrate the effect of the target size ratio on the number of measurements M needed in terms of reconstruction MSE and PSNR (the target size ratio is expressed as a ratio between non-zero pixels representing the target and the total size of the image frame, which reveals how much space the target acquires and how sparse the image is). It is clear from Fig.7 that for smaller target sizes, lower values of M are used while at the same time achieving the least MSE and PSNR of $\approx 33dB$ as in Fig.8(a) and 8(b), respectively. While for larger target sizes, a higher M is required to achieve the same performance achieved for frames with smaller targets. Experiments were carried out using the same M set to 50 for the 2 schemes (different sparsity levels). For example, frames with small size targets gave better reconstruction results in terms of least MSE and a $33dB$ PSNR as in Fig.8(a) and 8(b). Whereas, if the targets size grew bigger such as acquiring 60% space of the total frame size, with M set constant reconstruction results in high MSE and only $18dB$ PSNR. In that case M should be set to 90 or higher based on the sparsity nature to reach a low MSE and a PSNR of $\approx 30dB$ that was attained by lower M ($M = 50$) when compressing frames with targets of size $< 10\%$ of the frame size. These results reflect the constraint of the lower bound of M discussed in sec.II and give a key to the problem when M is required to be kept as small as possible. Where in that case the size of targets is controlled by zooming or changing the location of sensor nodes while bearing in mind to keep the scene of interest in the camera's field of view. By taking snapshots from a further location the total space acquired by the target is hence reduced and as a result M can be reduced, and the goal of reducing the size of transmitted data is met .

Another performance indicator is the correlation coefficient. After reconstructing the compressed measurements, the correlation coefficient indicates how likely the reconstructed frame correlates with the original one. Fig.9 shows by increasing M till reaching its optimum

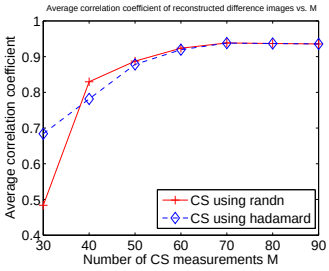


Fig. 9. Correlation coefficient for different M

values the correlation coefficients is nearly 100%, this implies that CS has not affected the image quality after recovery, whereas less number of measurements were required reducing the size of transmitted data.

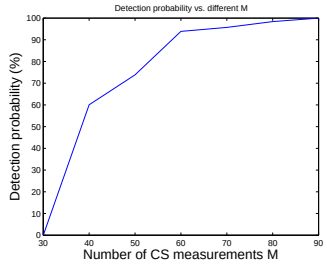


Fig. 10. Probability of detection vs different values of M

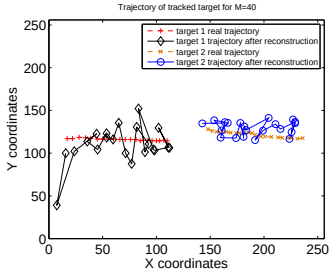
Fig.10 shows the probability of detection for different values of measurements M, it is clear from the graph that for lower values of M the target is misdetected. This reflects the fact that the reconstruction can not be guaranteed with lower values of M. The probability of detection increases till reaching 100% as M increases to its optimum value selected during the CS process.

CS states that when enough measurements are used for compression, the reconstruction is done with high accuracy depending on a lower bound of M. Trajectory tracking of moving targets is considered to reflects the degree of reconstruction accuracy. Tracking reliability is tested by comparing the moving target's real and predicted trajectories using LMS. Fig.11 and 12 show the (x,y) position plots of the path tracked for the targets in the camera's scene. Fig.11(a) and 11(b) show that (for "outdoor scheme") for lower values of $M <$ optimum value (40 and 70 respectively), frames can not be reconstructed properly and as a result the targets tracks are not matching their real trajectories, whereas for optimum values of M reaching 90, LMS accurately predicted the target's locations and the results are closely matching the real target trajectory before compression. Fig.12 illustrates the same for "indoor scheme".

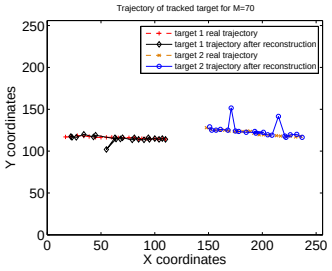
VI. Conclusion

Experiments were carried out to evaluate the performance of CS and its effect on target detection and tracking. Simulations have shown that CS is a strong candidate to reduce the size of images as WVSNs are resource constrained (Limited storage, channel bandwidth). Results have shown that using CS up to 31% measurements of data are required to be transmitted, while preserving the reconstruction quality which is measured in terms of MSE and PSNR. The reconstruction MSE decreases till reaching the lower bound on the number of compressed measurements while preserving the acceptable PSNR. In addition, for different schemes where the sparsity nature of each image differs, CS chooses the compression rates accordingly. Moreover, surveillance application within WVSNs is one of the important applications that requires high detection reliability and

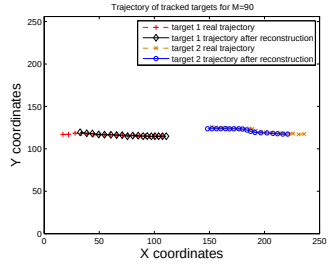
ISBN: 978-1-61804-236-1



(a) M=40

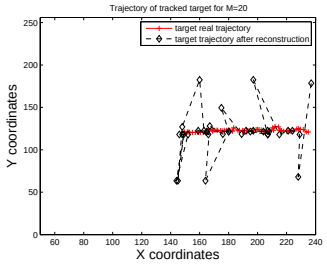


(b) M=70

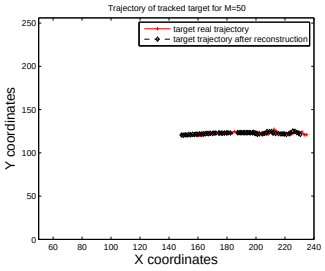


(c) M=90

Fig. 11. Comparing predicted trajectory of multi-targets using LMS for "outdoor scheme" (using different M for CS)



(a) M=20



(b) M=50

Fig. 12. Comparing predicted trajectory of single target using LMS for "indoor scheme" (using different M for CS)

robust tracking. Hence, CS should not affect the performance of target tracking. After image reconstruction, the impact of CS on target tracking is investigated using LMS to predict target's next location. Target's trajectory tracking has been used as a performance indicator for the LMS algorithm. Results have demonstrated that the predicted path closely matches the target's real path which illustrates the accuracy of LMS and that CS has not affected the performance of target detection and tracking.

REFERENCES

[1] A.Sharif, V.Potdar, and E.Chang, "Wireless multimedia sensor network technology: A survey," in *Proceedings of Industrial Informatics, 7th IEEE International Conference, 2009. INDIN 2009.*, June 2009, pp. 606 –613.

[2] I.F.Akyildiz, T.Melodia, and K.R.Chowdhury, "A survey on wireless multimedia sensor networks," *computer Networks*, vol. 51, pp. 921–960, March 2007.

[3] S.Soro and W.Heinzelman, "A survey on visual sensor networks," *Hindawi publishing corporation, Advances in Multimedia*, vol. 2009, no. 640386, pp. 1–21, May 2009.

[4] Y.Charfi, B.Canada, N.Wakamiya, and M.Murata, "Challenges issues in visual sensor networks," in *IEEE on wireless Communications*, April 2009, pp. 44–49.

[5] X.Wang, S.Wang, and D.Bi, "Distributed visual-target-surveillance system in wireless sensor networks," *IEEE Transactions on Systems, MAN, and Cybernetics*, vol. 39, no. 5, pp. 1134–1146, October 2009.

[6] X.Wang, S.Wang, D.W.Bi, and J.J.Ma, "Distributed peer-to-peer target tracking in wireless sensor networks," *MDPI, open access journal on the science and technology of sensors and biosensors*, vol. 7, pp. 1001–1027, 2007.

[7] X.Wang and S.Wang, "Collaborative signal processing for target tracking in distributed wireless sensor networks," *Elsevier journal on Parallel and distributed computing*, vol. 67, p. 501 515, 2007.

[8] D. Simon, "Kalman filtering with state constraints: a survey of linear and nonlinear algorithms," *The Institution of Engineering and Technology, Control theory applications*, vol. 4, no. 8, pp. 1303–1318, 2010.

[9] J.C.Noyer, P.Lanvin, and M.Benjelloun, "Non-linear matched filtering for object detection and tracking," *Elsevier Pattern Recognition Letters*, vol. 25, pp. 655–668, 2004.

[10] S. Lefevre and N. Vincent, "Real time multiple object tracking based on active contours," September 2004.

[11] J.Malcolm, Y.Rathi, and A.Tannenbaum, "Multi-object tracking through clutter using graph cuts," in *The International Conference on Computer Vision (ICCV)*, 2007.

[12] S.Haykin, *Adaptive Filter Theory*. Prentice Hall, 2002, vol. 0-13-048434-2, ch. Least mean square adaptive filters, pp. 231–247.

[13] P.S.R.Diniz, *Adaptive Filtering*. The Springer International Series in Engineering and Computer Science, January 1997, vol. 399, ch. The Least-Mean-Square (LMS) Algorithm, pp. 79–135.

[14] R. Baraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, pp. 118–124, July 2007.

[15] V.Cevher, A.Sankaranarayanan, M. Duarte, D.Reddy, R. Baraniuk, and R.Chellappa, "Compressive sensing for background subtraction," 2008.

[16] E. Wang, J. Silva, and L. Carin, "Compressive particle filtering for target tracking," in *IEEE/SP 15th Workshop on Statistical Signal Processing, SSP*, September 2009, pp. 233 –236.

[17] J.Romberg, "Imaging via compressive sampling," *IEEE Signal Processing Magazine*, pp. 14–20, March 2008.
[18] A.Mahalanobis and R.Muise, "Object specific image reconstruction using a compressive sensing architecture for application in surveillance systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 45, no. 3, pp. 1167–1180, July 2009.
[19] E.J.Candes and M.B.Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, pp. 21–30, March 2008.
[20] E.J.Candes, "Compressive sampling," in *Proc. of the International Congress of Mathematicians*, 2006.
[21] R.Baraniuk, "Compressive sensing," *IEEE signal processing magazine*, pp. 118–124, 2007.
[22] A.Hormati, O.Roy, Y.M.Lu, and M.Vetterli, "Distributed sampling of signals linked by sparse filtering: theory and applications," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1095–1109, March 2010.
[23] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, December 2007.
[24] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
[25] F. Cheng and Y. Chen, "Real time multiple objects tracking and identification based on discrete wavelet transform," *Elsevier Pattern Recognition Journal*, vol. 39, p. 1126 1139, 2006.
[26] "Caviar datasets," Dataset: EC Funded CAVIAR project/IST 2001 37540, http://homepages.inf.ed.ac.uk/rbf/CAVIAR/, 2001.
[27] H. Jiang, W. Deng, and Z. Shen, "Surveillance video processing using compressive sensing," *arXiv preprint arXiv:1302.1942*, 2013.

One-dimensional cutting stock model for joinery manufacturing

Ivan C. Mustakerov and Daniela I. Borissova

***Abstract*—**The current paper describes one-dimensional cutting stock model for joinery manufacturing. The joinery elements differ in size and number that are specific for each particular project. The goal is to determine the optimal length of blanks (which are usually ordered with equal size in large quantities) in order to satisfy the demand for all joinery elements. Along with this, it is necessary to find the optimal cutting patterns that minimize the overall trim waste. For the goal, one-dimensional cutting stock model for joinery manufacturing using combinatorial optimization is proposed. Numerical example of real-life problem is presented to illustrate the applicability of the proposed approach.

***Keywords*—**Combinatorial optimization model, joinery manufacturing, linear programming model, one dimensional cutting stock problem.

I. Introduction

THE cutting-stock problem has many applications in industry. This problem arises when the available material has to be cut to fulfill certain goals as cutting patterns with minimal material waste and cost efficient production, higher customer satisfaction, etc. In general, cutting stock problems consist in cutting large pieces (*blanks*), available in stock, into a set of smaller pieces (*elements*) accordingly to the given requirements, while optimizing a certain objective function. These problems are relevant in the production planning of many industries such as the metallurgy, plastics, paper, glass, furniture, textile, joinery manufacturing, etc. In the last four decades cutting stock problems have been studied by an increasing number of researchers [1]-[5]. The interest in these problems is provoked by the many practical applications and the challenge they provide to researchers. On the first glance they are simple to formulate, but in the same time they are computationally difficult to solve. It could be summarized that: cutting and packing problems [6] belong to the class of NP-hard problems; solution of these problems extensively uses mathematical programming and combinatorial methods; many real-life problems are computationally hard and can be formalized only as NP-hard problems. The continuous growth of the prices of the materials and of the energy requires minimization of the production expenses for every element.

I. C. Mustakerov is with the Institute of Information and Communication Technology at the Bulgarian Academy of Sciences, Sofia – 1113, Bulgaria, Department of Information Processes and Decision Support Systems (phone: 3952 9793241; e-mail: mustakerov@iit.bas.bg).

D. I. Borissova is with the Institute of Information and Communication Technologies at Bulgarian Academy of Sciences, Sofia – 1113, Bulgaria, Department of Information Processes and Decision Support Systems (phone: 3952 9792055; e-mail: dborissova@iit.bas.bg).

Most materials used in the industry are supplied of standard forms and lengths, and direct use of such forms is most cases are impossible. They should be cut in advance to some size, expected to be optimal in the sense of trim waste. This can be done using various methods of cutting planning. The problem of optimal cutting is that different size elements have to be manufactured using blanks of single standard size. This demands developing of methods for optimal cutting of source material. The one-dimensional cutting stock problem (1D-CSP) is one of the crucial issues in production systems, which involve cutting processes. The classical 1D-CSP addresses the problem of cutting stock materials of length in order to satisfy the demand of smaller pieces while minimizing the overall trim loss. Kantorovich first formulates 1D-CSP [7], [8] and Gilmore and Gomory [9], [10] propose the first solution methodology for the cutting stock problems.

In most cases, cutting stock problem is formulated as an integer linear programming optimization problem that minimizes the total waste while satisfying the given demand [11]. In [12] a review of some linear programming formulations for the 1D-CSP and bin packing problems, both for problems with identical and non-identical large objects, is presented. It is investigated haw different ways of defining the variables and structure of the models affect the solvability of problems. Because of NP-hard nature of cutting stock problems finding an optimal solution in reasonable time is essentially difficult and often researchers turn to heuristic algorithms to deal with this kind of complex and large-sized problem [3], [13]. Some researchers look for solutions of 1D-CSP in which the non-used material in the cutting patterns may be used in the future, if large enough [4] . A two-stage decomposition approach for 1D-CSP is proposed in [14]. In the first stage is performed calculation of the total number of patterns that will be cut and generation of the cutting patterns through a heuristic procedure. On the second stage optimal cutting plan is determined. In [15] a new approach to cutting stock problem is proposed where a ‘‘good’’ solution is seeking for consecutive time periods. It is adjusted to situations where useful stock remainders can be returned to the warehouse between time periods and used lately for other orders. A similar problem for wood industry is described in [16]. It is stated that cutting problems from the practice usually have its own specificity that do not allow the application of known models and solution algorithms. In many cases, proper modifications are needed or even completely new methods

ISBN: 978-1-61804-236-1

have to be developed on order to cope with real word requirements.

The current paper proposes new approach for optimization of real-life 1D-CSP from the joinery manufacturing practice. A combinatorial optimization task is formulated to determine the optimal length of the blanks and optimal cutting patterns in sense of minimal waste. In contrast to other 1D-CSPs, the optimal length of the blanks and optimal cutting patterns are defined simultaneously as a result of single optimization task solution. A proper algorithm for practical application of the proposed approach is defined and numerically tested using real-life data.

II. Problem Description

Aluminum or PVC blanks usually are supplied from the factory with fixed length of 6 meters. These blanks are used to cut out different elements of joinery. The joinery elements differ in size and number that are specific for each particular project. The goal is to determine the optimal length of blanks (which are usually ordered with equal size in large quantities) in order to satisfy the demand for all joinery elements. Along with this, it is necessary to find the optimal cutting patterns minimizing the waste. In [17] an in-depth investigation of joinery modules used in a wide range of buildings is performed. It was found that the number of joinery types in the apartments could be reduced to a certain number of unified modules. For example, in case of a middle size flat, these modules involve four modules:

- Module 1 is used for 4 doors with dimensions 2200 mm x 730 mm
- Module 2 is used for 2 doors with dimensions 2000 mm x 650 mm
- Module 3 is used for 1 window with dimensions 1400 mm x 1400 mm
- Module 4 is used for 2 windows with dimensions 1700 mm x 2100 mm

The investigated cutting stock problem can be narrowed down to definition of optimal length of blanks and optimal cutting patterns for modules used in an apartment. The problem can be described as follows: a factory has to fulfill order of blanks with certain length needed to assemble a given number of modules, consisting of elements with known length and number. For the sake of simplicity of the presentation only casement elements for the modules in the example above are summarized as a manufacturing order shown in Table I.

In practice, all PVC and aluminum profiles for doors and windows come with fixed length of 6 meters. However, this is not mandatory requirement and it is possible to order blanks with different length. There are no obstacles to order to the manufacturing company to produce a number of blanks with different length than standard 6 meters – for example any length between 5 and 7 meters. When the optimal length of blanks is determined, the next step is to define the optimal cutting patterns of joinery elements for each blank.

TABLE I
Joinery Elements Length and Demand

Element j	Length l_i, mm	Demand k_{ij}
1	$l_1 = 650$	4
2	$l_2 = 730$	8
3	$l_3 = 1400$	4
4	$l_4 = 1700$	4
5	$l_5 = 2000$	4
6	$l_6 = 2100$	4
7	$l_7 = 2200$	8

The problem of optimal joinery manufacturing can be investigated as 1D-CSP by means of proper mathematical modeling.

III. Mathematical Model Formulation

The described one-dimensional cutting stock problem for joinery is formalized via combinatorial optimization model. In contrast to other similar models it allows determining optimal length of blanks and optimal cutting patterns minimizing the trim loss, accordingly given demands of joinery elements. This type of functionality of the model requires introducing of inequalities for each of blanks. That means there is a necessity of knowing in advance the number N of the blanks. Number N can be calculated as overall demand of joinery elements divided by the length L of the blanks. On the other hand, the length L of the blanks is to be determined after solution of the optimization task. This "recursive" type of problem can be overcome taking into account that length L will have some value close to the standard length of 6 meters. Having this in mind, number of blanks N can be calculated as overall demand of joinery elements divided by the length of 6 meters, rounded to integer value. Then this value of N can be used to formulate the optimization task as:

$$\min \rightarrow \sum_{i=1}^{N} (L - L_i), i = 1,..., N \tag{1}$$

subject to

$$\forall i : L_i = \sum_{j=1}^{J} x_{ij} l_j, \; j = 1,..., J \tag{2}$$

$$\forall i : L_i \le L \tag{3}$$

$$\forall j : \sum_{i=1}^{N} x_{ij} = k_{ij} \tag{4}$$

$$(6 - \Delta_{\min}) \le L \le (6 + \Delta_{\max}) \tag{5}$$

$$\forall j : \; x_{ij} = \begin{cases} \textit{binary integer } 0 \textit{ or } 1, \text{ if } N \le k_{ij} \\ \textit{integer}, \quad \text{otherwise} \end{cases} \tag{6}$$

where N is number of blanks; L is length of blanks; L_i is the utilized length of each blank; l_j is length of joinery elements; x_{ij} are decision variables assigned to each element for particular blank; k_{ij} represents the demand of each element.

The objective function (1) minimizes the sum of trim loss for each blank. The optimal cutting pattern for each of the blanks is defined by decision variables x_{ij} in (2). Depending on the given particular joinery project, the decision variables (6) could be binary integer variables or integer variables. For example, if the number of the blanks is less than the maximum demand of some element, then the decision variables x_{ij} are to be considered as integers. This statement allows the model to allocate more than 1 element within cutting pattern in the blank to satisfy the elements demand. This elements demand is satisfied by (4). Deviation from the standard length of 6 meters is represented by Δ_{min} and Δ_{max} both approximately in the range of 1 meter.

IV. Numerical Illustration

The demand of elements for the example of joinery manufacturing order from Table I is illustrated in Fig. 1.

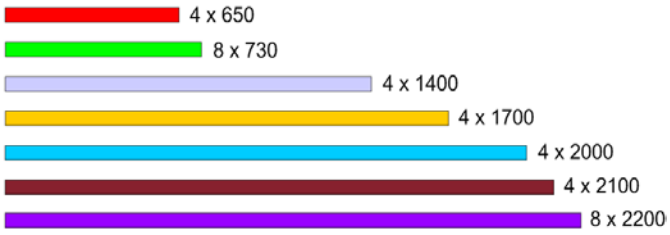


Fig. 1. Joinery elements and demand

Using the input data from Table I the following steps are performed:

1) Determination of total length of all elements considering their demand L_{sum} = 54840 mm;

2) Determination number of blanks N as rounded to integer result of the total elements length 54840 mm divided by 6000 mm as 54840/6000 = 9.14 => N = 9 and setting of deviations $\Delta_{min} = \Delta_{max}$ = 1000 mm.

3) Formulation of optimization task:

$$\min \{(L - L_1) + (L - L_2) + (L - L_3) + (L - L_4) + (L - L_5) + (L - L_6) + (L - L_7) + (L - L_8) + (L - L_9)\} \quad (7)$$

subject to

$$x_{11}l_1 + x_{12}l_2 + x_{13}l_3 + x_{14}l_4 + x_{15}l_5 + x_{16}l_6 + x_{17}l_7 = L_1 \quad (8a)$$
$$x_{21}l_1 + x_{22}l_2 + x_{23}l_3 + x_{24}l_4 + x_{25}l_5 + x_{26}l_6 + x_{27}l_7 = L_2 \quad (8b)$$
$$x_{31}l_1 + x_{32}l_2 + x_{33}l_3 + x_{34}l_4 + x_{35}l_5 + x_{36}l_6 + x_{37}l_7 = L_3 \quad (8c)$$
$$x_{41}l_1 + x_{42}l_2 + x_{43}l_3 + x_{44}l_4 + x_{45}l_5 + x_{46}l_6 + x_{47}l_7 = L_4 \quad (8d)$$
$$x_{51}l_1 + x_{52}l_2 + x_{53}l_3 + x_{54}l_4 + x_{55}l_5 + x_{56}l_6 + x_{57}l_7 = L_5 \quad (8e)$$
$$x_{61}l_1 + x_{62}l_2 + x_{63}l_3 + x_{64}l_4 + x_{65}l_5 + x_{66}l_6 + x_{67}l_7 = L_6 \quad (8f)$$
$$x_{71}l_1 + x_{72}l_2 + x_{73}l_3 + x_{74}l_4 + x_{75}l_5 + x_{76}l_6 + x_{77}l_7 = L_7 \quad (8g)$$
$$x_{81}l_1 + x_{82}l_2 + x_{83}l_3 + x_{84}l_4 + x_{85}l_5 + x_{86}l_6 + x_{87}l_7 = L_8 \quad (8h)$$
$$x_{91}l_1 + x_{92}l_2 + x_{93}l_3 + x_{94}l_4 + x_{95}l_5 + x_{96}l_6 + x_{97}l_7 = L_9 \quad (8i)$$

$$L_1 \le L \quad (9a)$$
$$L_2 \le L \quad (9b)$$
$$L_3 \le L \quad (9c)$$
$$L_4 \le L \quad (9d)$$
$$L_5 \le L \quad (9e)$$
$$L_6 \le L \quad (9f)$$
$$L_7 \le L \quad (9g)$$
$$L_8 \le L \quad (9h)$$
$$L_9 \le L \quad (9i)$$

$$x_{11} + x_{21} + x_{31} + x_{41} + x_{51} + x_{61} + x_{71} + x_{81} + x_{91} = 4 \quad (10a)$$
$$x_{12} + x_{22} + x_{32} + x_{42} + x_{52} + x_{62} + x_{72} + x_{82} + x_{92} = 8 \quad (10b)$$
$$x_{13} + x_{23} + x_{33} + x_{43} + x_{53} + x_{63} + x_{73} + x_{83} + x_{93} = 4 \quad (10c)$$
$$x_{14} + x_{24} + x_{34} + x_{44} + x_{54} + x_{64} + x_{74} + x_{84} + x_{94} = 4 \quad (10d)$$
$$x_{15} + x_{25} + x_{35} + x_{45} + x_{55} + x_{65} + x_{75} + x_{85} + x_{95} = 4 \quad (10e)$$
$$x_{16} + x_{26} + x_{36} + x_{46} + x_{56} + x_{66} + x_{76} + x_{86} + x_{96} = 4 \quad (10f)$$
$$x_{17} + x_{27} + x_{37} + x_{47} + x_{57} + x_{67} + x_{77} + x_{87} + x_{97} = 8 \quad (10g)$$

$$5 \le L \le 7 \quad (11)$$

$$x_{ij} \text{ – binary integer: 0 or 1} \quad (12)$$

The relations (8) in combination with inequalities (9) define optimal cutting patterns for each particular blank. The optimal cutting patterns are defined not to exceed the length of the blanks and to satisfy the requested demand of elements expressed by (10). The objective function (7) seeks for solution that minimizes the waste of all blanks. The optimal length of blanks is to be defined within interval of 5 to 7 meters (11). In this example the decision variables for optimal cutting patterns are binary integer variables (12).

The solution the formulated mixed integer optimization task (7) – (12) determines the optimal length of blanks; total waste; waste for each blank; and used length of each blank, as shown in Table II.

TABLE II
Optimal Solution Results

Optimal length of blanks L, mm	Total waste for order, mm	Used length of each blank, mm	Waste for each blank, mm
6550	4110	L_1 = 6330	220
		L_2 = 6330	220
		L_3 = 6030	520
		L_4 = 6030	520
		L_5 = 5680	870
		L_6 = 5680	870
		L_7 = 5680	870
		L_8 = 6530	20
		L_9 = 6550	0

ISBN: 978-1-61804-236-1

The optimal cutting patterns defined by the values of the binary integer variables for each blank are shown in Table III.

TABLE III
OPTIMAL CUTTING PATTERNS FOR EACH BLANK

	Element1	Element2	Element3	Element4	Element5	Element6	Element7
L_1	0	1	1	0	1	0	1
L_2	0	1	1	0	1	0	1
L_3	0	1	1	1	0	0	1
L_4	0	1	1	1	0	0	1
L_5	1	1	0	0	0	1	1
L_6	1	1	0	0	0	1	1
L_7	1	1	0	0	0	1	1
L_8	0	1	0	1	1	1	0
L_9	1	0	0	1	1	0	1

V. RESULT ANALYSIS AND DISCUSSION

The defined optimal length of blanks to fulfill the order is 6550 mm and the overall minimum waste is 4110 mm. The graphical illustration of optimal cutting patterns for each of the blanks is shown in Fig. 2.

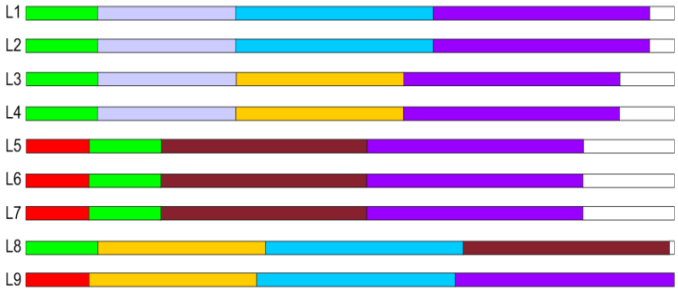
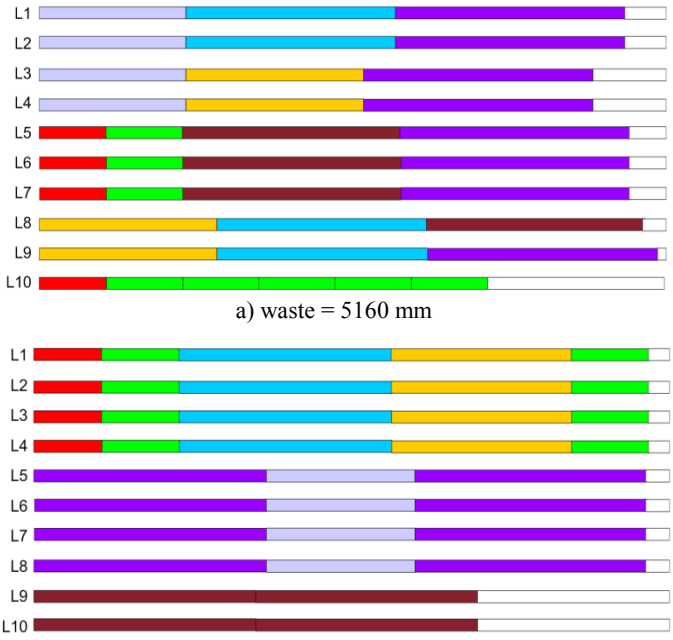


Fig. 2. Optimal cutting patterns for blanks
(L = 6550 mm, waste = 4110 mm)

It is compared with cutting patterns combinations defined by experienced practitioners for standard length of blanks equal to 6000 mm. The comparison shows that without optimization the trim loss is bigger as shown on Fig. 3.

The proposed optimization approach determines the optimal length of blanks that is increased toward standard length with 550 mm. This reduces number of needed blanks to fulfill the requested order and waste and costs as compared to the case of standard length using. Using of standard length of 6 m not only increases the trim loss but also increases the number of required blanks to execute the order. That is important for large joinery work projects in means of increasing of transportation costs.

Due to NP-hard nature of considered problems, the computational time increases essentially with increasing the number of decision variables. The formulated mixed integer linear optimization task (7) – (12) is solved by LINGO solver using branch-and-bound method [18].



a) waste = 5160 mm

b) waste = 5160 mm

Fig. 3. Cutting patterns for standard blank length L = 6000 mm

The solution time for the described example with 64 integer variables amounts to 1 hour, 23 minutes and 50 seconds on PC with 2.93 GHz Intel i3 CPU and 4 GB RAM. The task solution report is shown in Fig. 4.

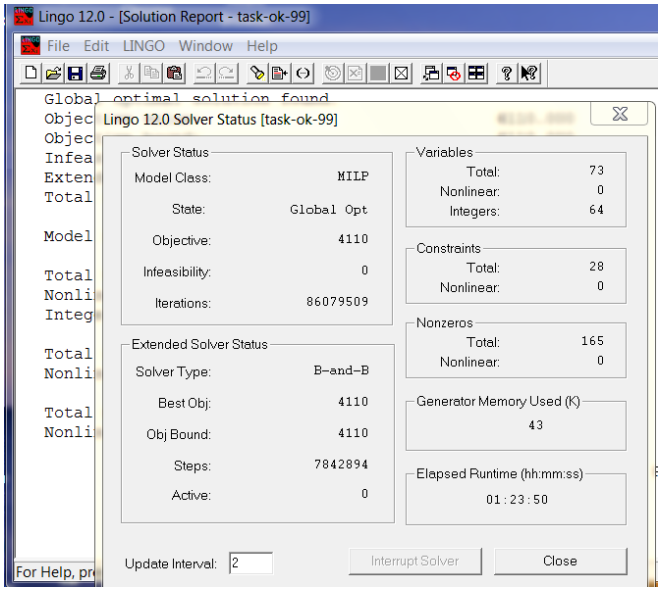


Fig. 4. Task solution report

VI. CONCLUSION

In the paper, joinery work manufacturing problem is investigated as one-dimensional cutting stock problem by means of combinatorial optimization. The advantage of the proposed approach is the possibility to determine simultaneously the optimal length of the blanks and optimal cutting patterns for each blank. In contrast to heuristic approaches to this type of problems the described approach

defines solution as a global optimum. The reduction of cutting trim loss is one of the main problems in joinery manufacturing. This problem turns to be important especially for large scale projects where the joinery work for a whole building or for several buildings has to be done. The described approach can contribute not only to reduce the trim loss via optimization of length of the blanks and cutting patterns, but also could decrease the overall production time and costs.

Future investigations are to be done with different large scale problems to determine the computational difficulties. Implementation of the proposed approach in a software tool for joinery work design will help the practitioners to reduce costs and will contribute to their competitiveness.

ACKNOWLEDGMENT

The research work reported in the paper is partly supported by the project AComIn "*Advanced Computing for Innovation*", grant 316087, funded by the FP7 Capacity Programme (Research Potential of Convergence Regions).

REFERENCES

[1] A. Mobasher and A. Ekici, "Solution approaches for the cutting stock problem with setup cost". *Computers & Operations Research*, vol. 40, 2013, pp. 225-235.

[2] A. C. Dikili, E. Sarioz and N. A. Pek, "A successive elimination method for one-dimensional stock cutting problems in ship production". *Ocean Engineering*, vol. 34, 2007, pp. 1841-1849.

[3] Y. Cui and Y. Lu, "Heuristic algorithm for a cutting stock problem in the steel bridge construction". *Computers & Operations Research*, vol. 36, 2009, pp. 612-622.

[4] C. Cherri, M. N. Arenales and H. H. Yanasse, "The one-dimensional cutting stock problem with usable leftover – A heuristic approach". *European Journal of Operational Research*, vol. 196, 2009, pp. 897-908.

[5] A. C. Dikili, A. C. Takinaci and N. A. Pek, "A new heuristic approach to one-dimensional stock-cutting problems", *Ocean Engineering*, vol. 35, no.7, 2008, pp. 637-645.

[6] E. A. Mukhacheva and A. S. Mukhacheva. "L. V. Kantorovich and Cutting-packing problems: New approaches to combinatorial problems of linear cutting and rectangular packing". *Journal of Mathematical Sciences*, vol. 133, no. 4, 2006, pp. 1504-1512.

[7] L. V. Kantorovich, Mathematical methods of organizing and planningproduction. *Management Science,* vol. 6, 1960, pp. 366-422.

[8] L. V. Kantorovich and V. A. Zalgaller, *Rational Cutting of Stock* [in Russian], Nauka, Novosibirsk, 1971.

[9] P. Gilmore and R. Gomory, "A linear programming approach to the cutting stock problem". *Operations Research,* vol. 9, no. 6, 1961, pp. 848-859.

[10] P. Gilmore and R. Gomory, "A linear programming approach to the cutting stock problem, part II". *Operations Research*, vol. 11, 1963, pp. 863-888.

[11] S. M. A. Suliman, "Pattern generating procedure for the cutting stock problem". *Int. Journal of Production Economics*, vol. 74, 2001, pp. 293-301.

[12] J. M. Valerio de Carvalho, "LP models for bin packing and cutting stock problems". *European Journal of Operational Research*, vol. 141, 2002, pp. 253-273.

[13] M. HMA Jahromi, R. Tavakkoli-Moghaddam, A. Makui and A. Shamsi, "Solving an one-dimensional cutting stock problem by simulated annealing and tabu search". *Journal of Industrial Engineering International*, vol. 8, no. 24, 2012, doi:10.1186/2251-712X-8-24.

[14] T. Aktin and R. G. Ozdemir. "An integrated approach to the one-dimensional cutting stock problem in coronary stent manufacturing". *European Journal of Operational Research*, vol. 196, 2009, pp. 737-743.

[15] P. Trkman and M. Gradisar. "One-dimensional cutting stock optimization in consecutive time periods", *European Journal of Operational Research*, vol. 179 , 2007, pp. 291-301.

[16] S. Koch, S. Konig and G. Wascher. "Linear Programming for a Cutting Problem in the Wood Processing Industry – A Case Study". *FEMM Working Paper* No 14, 2008.

[17] Ch. Korsemov, Hr. Toshev, I. Mustakerov, D. Borissova and V. Grigorova. "An optimal approach to design of joinery for renovation of panel buildings". *International Journal of Science and Engineering Investigations*, vol. 2, no. 18, 2013, pp. 123-128.

[18] Lindo Systems ver. 12, http://www.lindo.com

The performance of the MATLAB Parallel Computing Toolbox in specific problems

Dimitris N. Varsamis*, Christos Talagkozis*, Paris A. Mastorocostas*, Evangelos Outsios* and Nicholas P. Karampetakis†
*Department of Informatics Engineering
Technological Educational Institute of Central Macedonia, Serres 62124, Greece
Email: dvarsam@teiser.gr, talagozis@hotmail.com, mast@teiser.gr, outsios@teiser.gr
†Department of Mathematics
Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
Email: karampet@math.auth.gr

***Abstract*—In this work, we present the performance of three different parallel computing approaches of the MATLAB Parallel Computing Toolbox. In particular, we use the command "parfor", the command "spmd" and the technique "scheduler". The comparison of the three approaches in terms of computations and memory are presented. The three approaches are applied to two specific problems: a) searching of a value into a matrix and b) prime factorization. The first problem is bounded by MATLAB for the size of matrix, namely, has memory problems, and the second problem is bounded by MATLAB for numerical precision and time complexity. Finally, the executions of the corresponding parallel algorithms in a multi-worker lab are presented.**

***Index Terms*—MATLAB, Parallel Computing Toolbox, Searching, Prime factorization.**

I. Introduction

THE MATLAB Parallel Computing Toolbox (PCT) supports the two known parallel computer memory architectures such as the shared memory architectures and the distributed memory architectures [1], [2], [3], [4]. The shared memory architecture is implemented with the option "local" in a local computer with many cores [5]. Presently, the most common personal computers have CPU with two or four cores. The distributed memory architecture is implemented with the option "jobmanager" in a network of computers.
Additionally, the Matlab PCT supports the following parallel programming models: Data parallel, Distributed memory (message passing), Single Program Multiple Data (SPMD) and Multiple Program Multiple Data (MPMD). The command **parfor** implements the data parallel programming, the function **spmd** implements the SPMD and the technique **scheduler** implements both the distributed memory programming and the MPMD.
For the measurement of the performance of the Matlab PCT we select two specific problems: a) searching of a value into a matrix and b) prime factorization. The problem of the searching of a value into a matrix has limitation to the size of the matrix while the problem of the prime factorization has limitation to numerical precision and high computational cost.

This work was supported by the Research Committee of the Serres Institute of Education and Technology

A. The searching of a value into a matrix

The aim of this problem is to find a value into a large or a very large matrix. The questions for this problem are the following

- what is the maximum size of the matrix?
- how we can reduce the execution time?

The maximum size of a matrix in Matlab depends on the memory of the machine, the operating system and the release of Matlab. For example, in a PC with 4GB RAM, Windows XP(32 bit) and Matlab 7.3 the maximum memory which is used for an array is approximately 707 MB or 7.409×10^8 bytes (the function **memory** returns the maximum size of an array in Matlab), that means a matrix with

$$\frac{7.409 \times 10^8 \text{ bytes}}{8 \text{ bytes (double)}} = 92612500 \text{ cells}$$

Thus, we have a limit size for the simple use of Matlab.
Generally, if we consider M the maximum size of an array then the array has

$$\text{Total Cells} = \frac{\text{Maximum size}}{\text{bytes for double}} = \frac{M}{8}$$

Consequently, we use the PCT for the confronting of the limitations for the size of matrix and for the reduction of the execution time.
The serial approach in MATLAB of this problem is the following function:

```
function pos = search_for(x,key)
n=length(x);
pos=[];
for i=1:n
    if x(i)==key
        pos=[pos i];
    end
end
```

where `x` is an array (dataset), `key` is the searching value and `pos` is an array with the positions of the key value. In Table I the results of the performance tests for this function are presented. The size of the array x is bounded, thus we cannot use an array with size greater than of 2^{27} cells.

ISBN: 978-1-61804-236-1

TABLE I
THE PERFORMANCE OF THE SEARCHING PROBLEM IN SERIAL EXECUTION

Length of array x	Execution time	Comments
2^{15}	0.000333	
2^{20}	0.010854	
2^{25}	0.347254	
2^{27}	1.386923	Time problems
2^{28}	–	Memory problems

TABLE II
THE PERFORMANCE OF THE PRIME FACTORIZATION PROBLEM IN SERIAL EXECUTION

Order of number n	Execution time	Comments
2^{40}	0.069156	
2^{50}	2.189364	Time problems
2^{52}	6.152502	Time problems
2^{53}	–	Numerical precision problems
2^{54}	$\geq$ day	Use of Symbolic Toolbox

B. Prime factorization

The aim of this problem is to find the factorization of a very large integer number to two prime numbers. The prime factorization is the main idea of the RSA cryptosystem. The prime test is included in this problem, namely, an integer is or not prime number. The questions for this problem are the following

- what is the maximum number of digits of a number with double precision?
- how we can reduce the execution time?

The maximum number of decimal digits of a number with double precision is 16 because in double precision we have 8 bytes per number or $8 \times 8 = 64$ bits. That means a number with 53 binary digits or equivalently, a number with 16 decimal digits. This limitation can be confronted by using the Matlab Symbolic Toolbox. The main disadvantage of the symbolic toolbox is the high computational cost for the execution of simple arithmetic operations. Consequently, we use the PCT for the confronting of the limitations for the numerical precision and for the reduction of the execution time. The serial approach in MATLAB of this problem is the following function:

```
function [a,b] = prime_factorization(n)
k=double(floor(sqrt(sym(n))));
a=0;
b=0;
for m = 2:k
    if (mod(n,m) == 0)
        a=m;
        b=n/m;
        return
    end
end
```

where `n` is the number that is factorized and `a`, `b` are the factors, namely, $n = a \cdot b$. In Table II the results of the performance tests for this function are presented.

These specific problems are applied in three approaches of the Matlab PCT to reach conclusions and to find the advantages and disadvantages of each approach. In particular, in Section 2 we present the three parallel computing approaches which are the command **parfor** the command **spmd** and the technique **scheduler**. In Section 3 the performance of the parallel approaches are presented.

II. THE THREE PARALLEL COMPUTING APPROACHES

In this section, we present the command **parfor** which follows to Data parallel programming model. Additionally, we present the command **spmd** which follows to Simple Program Multiple Data programming model. Finally, we present the technique **scheduler** which follows to Distributed memory and Multiple Program Multiple Data programming model.

A. The command parfor

The command **parfor** (parallel loop) is MATLAB's simplest parallel programming command. **parfor** replaces the **for** command in cases where a loop can be parallelized. The loop iterations are divided up among different workers, executed in an unknown order, and the results gathered back to the main copy of MATLAB.

In the first problem we can replace the command **for** by the command **parfor** as see in the below function

```
function pos = search_parfor(x,key)
n=length(x);
pos=[];
parfor i=1:n
    if x(i)==key
        pos=[pos i];
    end
end
```

In the second problem we can not replace the command **for** by the command **parfor** because in body of loop the statement **return** exists which interrupts the execution of function when the condition of the statement if is true.

B. The command spmd

The **spmd** (single program multiple data) command is like working with a very simplified version of Message Passing Interface. There is one client process, supervising labs who cooperate on a single program. Each lab has an identifier, knows how many labs there are total, and can determine its behavior based on that identifier.

In the first problem we can partition the data according to the number of available labs and we can send these separately in each lab. The commands which are used are the following:

```
pos=0;
spmd(labs)
    for i=1:labs
        if(labindex==i)
           pos=search_for_file(key,n);
        end
    end
end
```

The function `search_for_file` is the same with the function `search_for` with the addition of a command

which it reads the data from a file.

Similarly, in the second problem we can partition the number of iterations according to the number of available labs and we can execute these separately in each lab. The commands which are used are the following:

```
k=double(floor(sqrt(sym(n))));
step=ceil(k/labs);
spmd(labs)
   for i=1:labs
     if(labindex==i)
         [a,b]=prime_factorization_spmd(n,...
                        1+step*(i-1),step*i);
     end
   end
end
```

The function `prime_factorization_spmd` is the following

```
function [a,b] = prime_factorization_spmd(n,x,y)
a=0;
b=0;
m=x;
while m<=y
    if (mod(n,m)==0 && m~=1)
        a=m;
        b=n/m;
        return
    end
    m=m+1;
end
```

In above function the arguments x and y are the lower and the upper index of iterations.

C. The technique scheduler

The technique **scheduler** with a Job Manager manage a big job which is divided into vary independent tasks. For simplicity, assume each task will be carried out by the same MATLAB function. Each task runs on a single processor (although there's no need to rule out parallelism) and has its own memory. Tasks do not communicate while running; they start with input, they return results upon completion. When all the tasks are completed, it is possible to gather, analyze and plot the combined results.

In the first problem we can apply the technique **scheduler** with the following commands

```
jm = findResource;
pj = createJob(jm);
set(pj,'MinimumNumberOfWorkers',1);
set(pj,'MaximumNumberOfWorkers',64);
set(pj,'FileDependencies',...
        {'dataset25.mat','search_for_file.m'});
for i=1:labs
    obj(i)=createTask(pj,...
          @search_for_file, 1, {key,log2(n)});
end
submit(pj);
waitForState(pj);
out=getAllOutputArguments(pj);
```

In above program we create task for each lab and we send the tasks with the corresponding files in each lab. The sending files are the function `search_for_file` with specific arguments `key`, `log2(n)` and the file `dataset25.mat` with the dataset.

In the second problem we can apply the technique **scheduler** with the following commands

```
k=double(floor(sqrt(sym(n))));
step=ceil(k/labs);
jm = findResource;
pj = createJob(jm);
set(pj,'MinimumNumberOfWorkers',1);
set(pj,'MaximumNumberOfWorkers',64);
set(pj,'FileDependencies',...
        {'prime_factorization_spmd'});
for i=1:labs
    obj(i)=createTask(pj,...
          @prime_factorization_spmd,...
          1, {n,1+step*(i-1),step*i});
end
submit(pj);
waitForState(pj);
out=getAllOutputArguments(pj);
destroy(pj);
```

In above program we create task for each lab and we send the tasks with the corresponding file in each lab. The sending files are the function `prime_factorization_spmd` with specific arguments `n,1+step*(i-1)` and `step*i`.

III. The performance of the parallel approaches

The performance tests are implemented in a lab with 24 computers. The specifications of each computer are Intel Core Quad CPU (Q9400) at 2600 GHz with 4 Gb RAM. We run the parallel programs with 1, 2, 4, 8, 16, 32 and 64 cores. The sizes of the dataset of the searching problem are 2^{20} and 2^{25}. The number of digits of the prime factorization problem are order of 2^{40}, 2^{50} and 2^{52}. In Tables III and IV the results of the performance test are presented.

For the searching problem we have the following characteristics:

- we cannot use a matrix with total size (number of cells) greater than $\frac{M}{8}$, therefore we do not gain in memory issues with command **parfor**.
- we can use a matrix with total size (number of cells) greater than $\frac{M}{8}$, therefore we gain in memory issues with the command **spmd** and the technique **scheduler**. For example, if the numbers of labs is 64 and the matrix size in each lab is 2^{26} then we have total size $64 \times 2^{26} = 2^{32}$.
- the execution times (see Table III) show us that we do not gain in execution time in relation to serial execution in all approaches.
- the execution times (see Table III) show us that in matrix with size of 2^{20} cells the command **parfor** has better times instead of **spmd** and **scheduler**.
- the execution times (see Table III) show us that in matrix with size of 2^{25} cells the command **spmd** has better times instead of **parfor** and **scheduler** and while the cores increasing the time is the same.

ISBN: 978-1-61804-236-1

TABLE III
EXECUTION TIMES FOR THE SEARCHING PROBLEM WITH MATRIX SIZE OF 2^{20} AND 2^{25}

cores	parfor	spmd	scheduler	parfor	spmd	scheduler
1	1.902	0.8367	1.1850	45.13	18.3763	23.4657
2	0.99	0.9235	1.2984	30.61	18.4263	30.3068
4	0.839	1.3980	1.5278	24.32	19.2775	41.0596
8	0.736	1.3143	1.8254	23.08	19.1012	52.6693
16	0.74	1.3900	2.4786	22.91	19.0348	72.3122
32		1.3630	3.6418		19.6694	109.2248
64		1.3431	6.1749		21.9759	187.8599

TABLE IV
EXECUTION TIMES FOR THE PRIME FACTORIZATION PROBLEM WITH NUMBER n ORDER OF 2^{40}, 2^{50} AND 2^{52}

cores	spmd	scheduler	spmd	scheduler	spmd	scheduler
1	0.2510	0.6341	4.0960	5.1087	10.8393	11.2883
2	0.1741	0.5984	2.0935	2.5601	5.6405	5.9604
4	0.1679	0.6360	1.2267	1.6284	3.2516	3.4032
8	0.1687	0.6317	0.6770	1.1801	1.6976	2.1890
16	0.1992	0.6297	0.4079	0.9058	0.9172	1.4083
32	0.2601	0.6659	0.3416	0.7943	0.5780	1.0249
64	0.4150	0.7617	0.4351	0.8650	0.5421	0.9388

For the prime factorization problem we have the following characteristics:

- In number with order of 2^{40} the execution times (see Table IV) are the same for the **spmd** and **scheduler**.
- In number with order of 2^{50} and 2^{52} the execution times (see Table IV) are decreasing while the cores are increasing for the **spmd** and **scheduler**.

The theoretical measurements of parallel computing are a) the speed up of the parallel algorithm which is given by

$$S_p = \frac{T_1}{T_p}$$

where T_1 is the execution time of the serial algorithm and T_p is the execution time of the parallel algorithm with p processors and b) the efficiency of the parallel algorithm which is given by

$$E_p = \frac{S_p}{p}$$

The efficiency of the searching problem are presented in Figures 1 and 2 and the efficiency of the prime factorization problem are presented in Figures 3,4 and 5.

IV. CONCLUSIONS

From the results of the performance tests and the theoretical measurements of parallel processing we conclude in the according observations: a) in problems with large or very large dataset we can use the MATLAB PCT so as to partition the dataset in each lab, which has its own memory. That is, we use the distributed memory parallel model. On the other hand, the PCT cannot reduce the execution times (very poor efficiency in all approaches). b) In problems with high computational cost we can use the PCT to reduce the computational time (good efficiency in very large numbers with small number of cores). In particular, the command **spmd** is the best choice for the searching problem. The command **spmd** is the best choice for the prime factorization problem with the number of cores less equal to 32 (efficiency $\geq 50\%$).

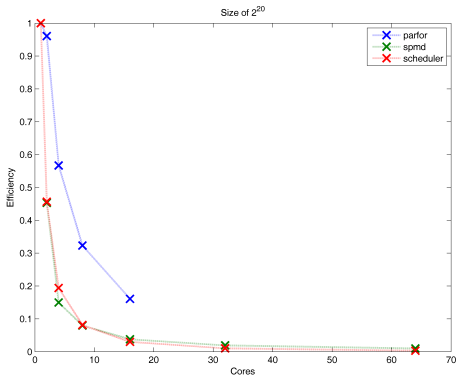


Fig. 1. The efficiency of the PCT in searching problem with matrix size of 2^{20}.

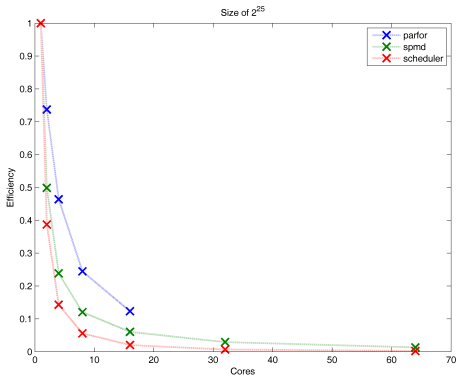


Fig. 2. The efficiency of the PCT in searching problem with matrix size of 2^{25}.

ISBN: 978-1-61804-236-1

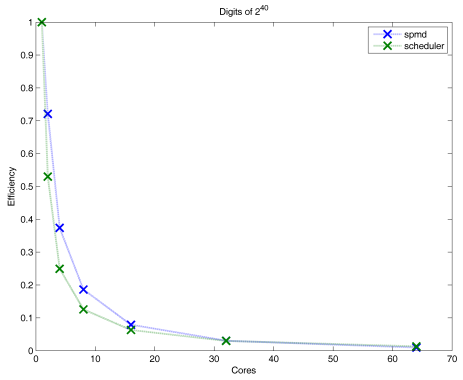


Fig. 3. The efficiency of the PCT in prime factorization problem with number with order of 2^{40}.

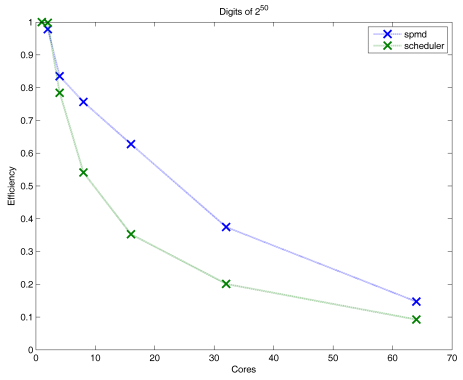


Fig. 4. The efficiency of the PCT in prime factorization problem with number with order of 2^{50}.

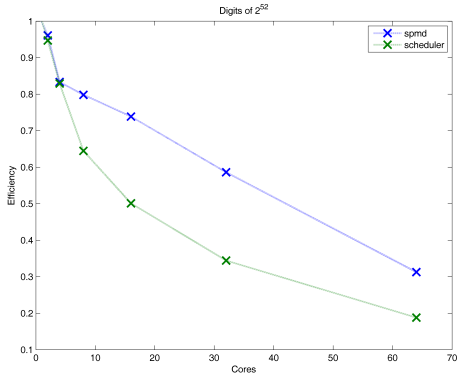


Fig. 5. The efficiency of the PCT in prime factorization problem with number with order of 2^{52}.

ACKNOWLEDGMENT

This work was supported by the Research Committee of the Serres Institute of Education and Technology.

REFERENCES

[1] J. Kepner, *Parallel MATLAB for Multicore and Multinode Computers*. Philadelphia, USA: SIAM, 2009.

[2] P. Luszczek, "Parallel programming in matlab," *International Journal of High Performance Computing Applications*, vol. 23, no. 3, pp. 277–283, 2009.

[3] C. Moler, "Parallel matlab: Multiple processors and multiple cores," *The MathWorks News & Notes*, 2007.

[4] G. Sharma and J. Martin, "Matlab : A language for parallel computing," *International Journal of Parallel Programming*, vol. 37, pp. 3–36, 2009.

[5] D. N. Varsamis, P. A. Mastorocostas, A. K. Papakonstantinou, and N. P. Karampetakis, "A parallel searching algorithm for the insetting procedure in matlab parallel toolbox," in *Federated Conference on Computer Science and Information Systems (FedCSIS), 2012*. IEEE, 2012, pp. 587–593.

[6] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Prentice Hall, 1989.

[7] J. W. Demmel, M. T. Heath, and H. A. van der Vorst, "Parallel numerical linear algebra," *Acta Numerica*, vol. 2, pp. 111–197, 1993.

[8] A. Grama, A. Gupta, G. Karypis, and V. Kumar, *Introduction to Parallel Computing*, 2nd ed. Addison-Wesley, 2003.

[9] C. Lin and L. Snyder, *Principles of Parallel Programming*. Boston, USA: Addison-Wesley, 2008.

[10] C. Moler, *Numerical Computing with MATLAB*, 2nd ed. SIAM, 2008.

An Approach to Development of Visual Modeling Toolkits

Alexander O. Sukhov, Lyudmila N. Lyadova

***Abstract* —** The approaches based on applying of metamodeling and domain-specific languages are widely used in software engineering. There are many different tools for creating visual domain-specific modeling languages with a possibility of determining user's graphical notations. However these tools possess disadvantages. The article presents an approach to the development of language workbench that allows to eliminate some restrictions of existing DSM-platforms. The MetaLanguage system is designed for creation of visual dynamic adaptable domain-specific modeling languages and for models construction with these languages. It allows executing transformations of the created models in various textual and graphical notations. Basic metalanguage constructions of this system are described. The formal description of modeling languages metamodel used in MetaLanguage is given. The architecture of MetaLanguage toolkit is presented.

***Keywords* —** DSM-platform, graphs, metamodeling, visual domain-specific languages.

I. Introduction

Creation of information systems with usage of the modern tools is based on the development of the various models describing the domain of information system, defining data structures and algorithms of system functioning. At implementation of *model-driven approach* to software development the models become a central element at all stages of system creation. The model-based approach is capable at information system creation to unite efforts of developers and domain experts. This approach makes the system more flexible, since for its change there is no necessity of modification of source code "by hand", it is enough to modify a visual model, and with this task even nonprofessional programmers can cope [1], [2].

At usage of this approach the models describing system from the various points of view, with a different level of abstraction and with usage of various modeling languages are created. For coordination 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 (object, process) characteristics that are important from the point of view of the modeling purpose. A model is created with usage of specific modeling language. A *metamodel* is a model of the language, which is used for models development, a *meta-metamodel (metalanguage)* is a language, on which metamodels are described.

For model-based approach implementation it is necessary to use toolkit, which will be convenient to various participants of system development process. The general-purpose modeling languages, such as UML, are not able to cope with this task, because they have some disadvantages [3], [4]:

- Diagrams are complicated for understanding not only for experts, who take part in system engineering, but in some cases even for professional developers.
- Object-oriented diagrams can not adequately represent domain concepts, since work is being done in terms of "class", "association", "aggregation", etc., rather than in domain terms.

That is why at implementation of model-based approach the domain-specific modeling languages (DSMLs), created to work in specific domains, are increasingly used. Domain-specific languages are more expressive, simple on applying and easy to understand for different categories of users as they operate with domain terms. For this reason now a large number of DSMLs is designed for creation of systems in different domains: artificial intelligence systems, distributed systems, mobile applications, real-time and embedded systems, simulation systems, etc. [5]–[7].

Despite of all DSMLs advantages they have one big disadvantage – complexity of their designing. If general purpose languages allow to create programs irrespectively to domain, in case of DSMLs for each domain, and in some cases for each task, it is necessary to create new domain-specific language. Another shortcoming of visual domain-specific language is that it is necessary to create convenient graphical editors to work with it.

To support the process of development and maintenance of DSMLs the special kind of software – *language workbench* (*DSM-platform*) is used [8]. Usage at DSMLs creation of a language workbench considerably simplifies the process of their designing. There are various DSM-platforms for creating visual DSMLs with the ability of determining user's graphical notation: MetaEdit+, Microsoft DSL Tools, Eclipse GMF, QReal, etc. Let's consider the most advanced language workbenches [9], [10].

This work was supported in part 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. Related Works

MetaEdit+ is a multiplatform language workbench that enables users to simultaneously work with several projects, each of which can have a few models [11], [12]. At usage of this DSM-platform besides a possibility of domain-specific language creation, the developer receives the CASE tool, into which this language is integrated. MetaEdit+ allows to use several DSMLs at system creation.

The approach based on metamodels (models of modeling languages) interpretation, instead of code generation, which is used in MetaEdit+ allows changing the DSMLs definition at run-time. The system allows working with languages and metalanguages universally, using the same tools. The disadvantage of MetaEdit+ is that this DSM-platform for export of models uses an own file format (MXT) and this affects the openness of technology.

Microsoft DSL Tools [13] and Eclipse GMF [14] technologies provide the user with advanced IDE MS Visual Studio and Eclipse respectively. Thanks to this there is a possibility of code completion on high-level languages "by hand", but it can lead to occasion of inconsistency of diagrams and source code.

Technology Eclipse GMF is most powerful of the above. However, its usage is impeded by high complexity, frequent releases of new versions and lack of documentation. In fact, Eclipse GMF is in a stage of intensive development.

Eclipse environment provides the user with tab GMF Dashboard, which allows to accelerate DSMLs development process by automatically generating some language components. On GMF Dashboard tab the sequence of the operations is represented. Tools of creation of plug-ins for Eclipse, which allow to build diagrams in current domain are realized with these operations.

The multiplatform system QReal [15] allows to define metamodels both in visual and textual view, therefore developers have a possibility to select the most suitable for them format of language description representation. Availability of an interpreter of behavioral diagrams and a debugger of the generated code puts this system to the same position as Microsoft DSL Tools, Eclipse GMF, which use for these purposes IDE. In QReal there is not a possibility of modification of DSMLs description at run-time.

Cases when DSMLs becomes part of other applications are common. For example, a specially designed language for describing business processes can be used in document circulation model. Therefore one more important characteristic of the DSM-platforms is the alienability of DSMLs from the development environment. Microsoft DSL Tools, Eclipse GMF are strongly associated with the development platforms – MS Visual Studio and Eclipse, respectively, therefore languages created by these workbenches can't be exported to external system.

The analysis of DSM-platforms has revealed some restrictions inherent in the majority of the considered systems:

1. Impossibility of multilevel modeling. Presence of such possibility would allow making changes at metalanguage description, to extend it with new constructions, thus bringing the metalanguage to the specifics of domain.
2. Modification of DSMLs description leads to necessity of regeneration of language editor: for modification DSMLs at first it is necessary to change its metamodel, to regenerate the source code of the editor, and only then it is possible to begin build models.
3. "Excess" functionality of the language workbench, which is not used at DSMLs creation. This functionality complicates the study of tools by the users, which are not professional programmers.
4. Lack of tools of horizontal models transformations. These means allow not only to create unified system description on the basis of the models constructed at various stages of system development, but also to generate source code according to user-specified template or to make conversion of the model described with one modeling language to model fulfilled in other graphical notation.

The MetaLanguage system eliminates some restrictions of the considered DSM-platforms.

III. MetaLanguage System

The DSM-platform MetaLanguage 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.

A. Metalanguage of MetaLanguage System

One of the basic elements of language workbench is the *metalanguage* (*meta-metamodel*), which is the language for describing of other languages. Thanks to presence of metalanguage the DSM-platform allows to create domain-specific languages for the various domains that operate with familiar for the user concepts. The main difference between metalanguage of MetaLanguage system from the MOF (Meta Object Facility) approach, used in the majority of DSM-platforms, is that thanks to interpretation of models at various abstraction levels, instead of the source code generation on their basis, it is possible to modify of DSML's constructions in dynamics, at models creation. Besides, the process of metamodel creation becomes multilevel, so developer defining metamodel and selecting it as the metalanguage, can use it to create other metamodels, and this process can be infinite.

The *basic elements* of the metalanguage of MetaLanguage system are entity, relationship and constraint.

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.

Visual language constructions in rare cases exist independently, more often they are in some way related to each other, therefore at metamodel creation importantly not only to define the basic language constructions, but also correctly specify the relationships between them. The *relationship* is used for describing a physical or conceptual links between

entities. Metamodel allows to create three types of relationships: *association*, *aggregation*, *inheritance*.

In practice quite often there are cases when it is necessary to impose some *constraints* on entities and relationships. Some of constraints are set by metamodel structure, and others are described on some languages. All constraints imposed on the metamodel in MetaLanguage system can be divided into two groups: constraints imposed on entities and constraints imposed on relationships.

Let's consider an example. A fragment of metamodel for UML Use Case diagrams is shown in Fig. 1. The metamodel contains two entities "Actor" and "Use Case".

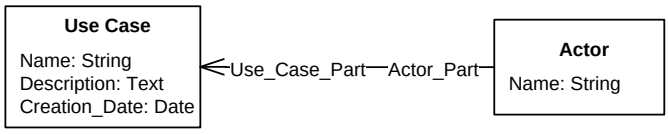


Fig. 1. Fragment of metamodel for UML Use Case diagrams

The entity "Use Case" has following attributes: "Name", "Description", "Creation_Date". The attribute "Name" has a string type and defines the "Use Case" name. The attribute "Description" sets the short description of the "Use Case". An attribute of entity "Actor" is a string attribute "Name", which specifies the name of the Actor.

B. Architecture of MetaLanguage System

The architecture of MetaLanguage is presented in Fig. 2. Uniform storage of all information about the system is the *repository*. It contains information about metamodels, models, entities, relationships, attributes, constraints. Information about the models and metamodels is stored uniformly, that allows to work with it by the same tool.

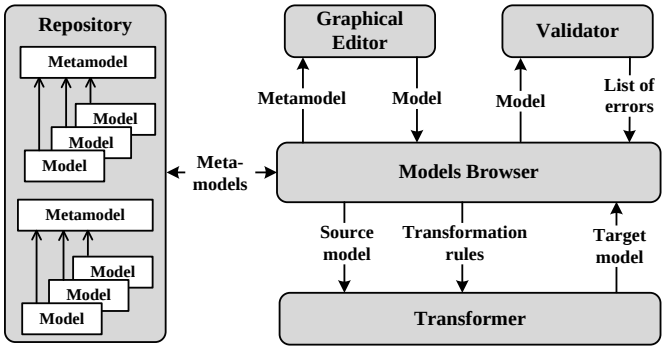


Fig. 2. Architecture of MetaLanguage system

The *browser of models* allows to load/save metamodels together with the models created on their basis, to fulfill over metamodels and models various operations (editing, constraint checking, transformation, etc.). The *graphical editor* is the component, which provides the user the tools for metamodels and models creation. The *validator* allows to check constraints specified by user at metamodel describing. The *transformer* is the component that provides the ability to fulfill horizontal transformations of models to text on target programming language or to visual models, described in other graphical notation.

Having described the basic components of a MetaLanguage system, let's consider how visual domain-specific modeling languages are designed with these tools.

Process of DSML definition begins with metamodel creation. For this purpose it is necessary to specify the main constructions of created language, to define relationships between them, to set constraints imposed on the metamodel entities and relationships. After building of metamodel the developer gets a customizable extensible visual modeling language.

Then the user can design models containing objects that describe specific domain concepts and links between them with using created DSML.

The validator should check up whether model correspond to constraints, which were imposed on metamodel elements.

Then the developer can save the constructed metamodels and models in the form of XML-files or transform these models to other textual or graphical notation.

At metamodel modification the system automatically makes all necessary changes in the models, which are created on the basis of this metamodel.

Using constructions "entity" and "relationship" it is possible to build any model, including an incorrect model in the current domain.

The metamodel of visual modeling language is a graph. There are several types of graphs that can be used for representation of visual languages: the classical graphs, digraphs, multigraphs, pseudographs, hi-graphs, hypergraphs, metagraphs and others [16]-[18].

As an analysis result of various types of graph it has been defined that the most appropriate formalism for describing the syntax of visual modeling languages in MetaLanguage system is pseudo-metagraph [19].

Metagraph is an ordered pair $G = (V, E)$, where V is a finite nonempty set of nodes, E is a set of edges. Each edge $e_k = (V_i, V_j), V_i, V_j \subseteq V$ connects two subsets of nodes.

Let's describe with usage of this formalism a metamodel of a visual modeling language.

IV. Formal Description of a Modeling Language Metamodel

Let $Ent = \{ent_i\}, i \in \aleph$ ($\aleph$ is a set of natural numbers) is a set of metamodel entities, number of set elements is potentially unlimited, but at every fixed point in time is finite.

The set of metamodel relationships denotes as $Rel = \{rel_i\}, i \in \aleph$, number of set elements is potentially unlimited, but at every fixed point in time is finite.

Let's introduce the following designations:

1) $EAttr_i = \{eattr_{i_j}\}, \overline{i = 1, |Ent|}, j \in \aleph$ is the set of metamodel graph nodes, which correspondence to entities attributes;
2) $RAttr_k = \{rattr_{k_l}\}, \overline{k = 1, |Rel|}, l \in \aleph$ is the set of metamodel graph nodes, which correspondence to relationships attributes;

ISBN: 978-1-61804-236-1

3) $ERest_i = \{erest_{i_j}\}, \overline{i=1,|Ent|}, j \in \aleph$ is the set of metamodel graph nodes, which correspondence to constraints imposed on entities;
4) $RRest_k = \{rrest_{k_l}\}, \overline{k=1,|Rel|}, l \in \aleph$ is the set of metamodel graph nodes, which correspondence to constraints imposed on relationships;
5) $EEA = \{eea_i\}, \overline{i=1,|Ent|}$ is the set of metamodel graph arcs connecting each entity with the set of its attributes;
6) $ERA = \{era_k\}, \overline{k=1,|Rel|}$ is the set of metamodel graph arcs connecting each relationship with the set of its attributes;
7) $EER = \{eer_i\}, \overline{i=1,|Ent|}$ is the set of metamodel graph arcs connecting each entity with the set of its constraints;
8) $ERR = \{err_k\}, \overline{k=1,|Rel|}$ is the set of metamodel graph arcs connecting each relationship with the set of its constraints;
9) $EERR = \{eerr_i\}, i \in \aleph$ is the set of arcs corresponding to links between entities and relationships.

The number of elements of sets $EAttr_i$, $RAttr_k$, $ERest_i$, $RRest_k$, EEA, ERA, EER, ERR, $EERR$ potentially is not limited, but it is finite at every fixed point in time.

The metamodel graph is the directed pseudo-metagraph $GMM = (V, E)$, where V is a nonempty set of graph nodes, E is set of graph arcs and these sets are defined by (1) and (2):

$$V = Ent \cup \left(\bigcup_{i=1}^{|Ent|} EAttr_i \right) \cup \left(\bigcup_{i=1}^{|Ent|} ERest_i \right) \cup \cup Rel \cup \left(\bigcup_{k=1}^{|Rel|} RAttr_k \right) \cup \left(\bigcup_{k=1}^{|Rel|} RRest_k \right) \quad (1)$$

$$E = EEA \cup EER \cup ERA \cup ERR \cup EERR \quad (2)$$

Let's consider an example. We will construct a metamodel graph for the entity "Use Case" of UML Use Case diagrams. Metamodel of this diagram type is shown in Fig. 1. Attributes of the entity "Use Case" are "Name", "Description", "Creation_Date", i.e. for given entity

$EAttr_i$ = {"Name", "Description", "Creation_Date"}.

The metamodel graph corresponding to the fragment of the "Use Case" entity is shown in Fig. 3. As can be seen from the figure:

$ERest_i = \varnothing$, $EEA = \{eea_i\}$, $EER = \varnothing$, $EERR = \varnothing$.

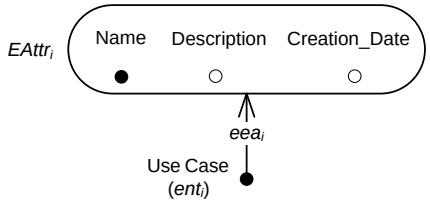


Fig. 3. Fragment of metamodel graph for "Use Case" entity

The model graph is defined similarly. The *model graph* is directed pseudo-metagraph $GM = (VI, EI)$ where VI is a nonempty set of graph nodes, EI is set of graph arcs and these sets are defined by (3) and (4):

$$VI = \bigcup_{i=1}^{|Ent|} \left(EntI_i \cup \left(\bigcup_{j=1}^{|EAttr_i|} EAttrI_{i_j} \right) \right) \cup \cup \left(\bigcup_{k=1}^{|Rel|} \left(RelI_k \cup \left(\bigcup_{l=1}^{|RAttr_k|} RAttrI_{k_l} \right) \right) \right) \quad (3)$$

$$EI = EEAI \cup ERAI \cup EERRI \cup T \quad (4)$$

In the graph model definition the following notation is used:

1) $EntI_i$ is the set of instances of i-th entity;
2) $EAttrI_{i_j}$ is the set of attribute values for j-th instance of i-th entity;
3) $RelI_k$ is the set of instances of k-th relationship;
4) $RAttrI_{k_l}$ is the set of attribute values for l-th instance of k-th relationship;
5) $EEAI$ is the set of arcs connecting each entity instance with set of attributes belonging to it;
6) $ERAI$ is the set of arcs connecting each relationship instance with set of attributes belonging to it;
7) $EERRI$ is the set of arcs corresponding to the links between entities instances and relationships instances;
8) T is the set of arcs of model graph, connecting instances of entities and relationships with those entities and relationships, on which basis they are created.

The number of elements of all these sets potentially is not limited, but it is finite at every fixed point in time.

On the basis of this mathematical model the operations of creation and interpretation of graph models were defined.

Actually these operations are *algorithms of vertical transformations of models* in forward and reverse direction. So at model creation the user, operating with metamodel entities and relationships, creates their instances, thus actually there is a mapping of metamodel graph to model graph. This mapping corresponds *to operation of graph model creation*. After model creation it is necessary to perform checking of the constraints, imposed on metamodel, and, in case of need, conversion of model description to other notation. At execution of these operations the system makes interpretation of model elements, i.e. defines with what entities and relationships they have been created. The mapping of model graph to metamodel graph is used for this purpose. This mapping corresponds to *operation of model graph interpretation*.

With usage of this formalism the mathematical model, which is a basis for implementation of the MetaLanguage system, has been constructed. According to the mathematical model and the requirements to this language workbench, the environment for visual DSMLs creation is designed and algorithms of the MetaLanguage functioning are developed: algorithms for creation/modification/removal of metamodels and models elements, algorithms of constraints checking, algorithms for vertical and horizontal models transformations.

ISBN: 978-1-61804-236-1

V. Creation of DSML with Usage of the MetaLanguage System

Let's consider an example: construct with usage of the MetaLanguage system the domain-specific language for creation of models of "Smart House" systems.

At first let's analyze the components, which can be a part of "Smart House" systems. The basic elements of systems of this type are:

- life-support systems: heating, air conditioning and ventilation, lighting, security;
- sensors (devices that are responsible for obtaining of various readings and their sending to central panel): motion, leakings, fire and a smoke, closing/opening of object;
- system management tools: voice control, remote control (from a remote computer, from phone, etc.), touch control (control by using of the touch screen of a central panel);
- central panel, which is responsible for receiving of data from sensors, management of life-support systems and obtaining of commands from the user.

For a unified description of entities corresponding to the different life-support systems, let's define the abstract entity "Life-support system", which has the following attributes: "Name", "Manufacturer", "Cost", "State" (defines the state of the system in current time). "Life-support system" entity has the following child entities: "Heating system", "Air conditioning and ventilation system", "Lighting system", "Security system" (see Fig. 4).

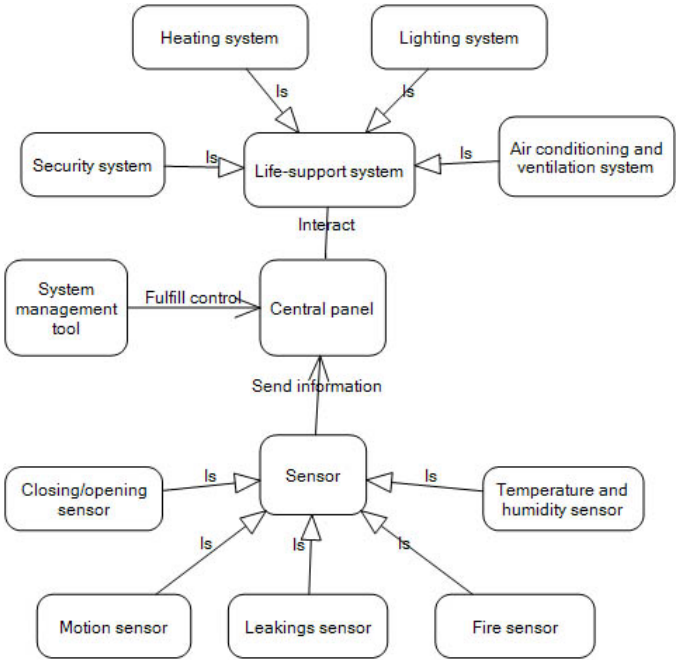


Fig. 4. Metamodel of visual DSML, created in MetaLanguage system

Entities "Heating system" and "Air conditioning and ventilation system" in addition to the inherited attributes have their own attribute "Temperature" containing value of temperature, which is necessary for supporting indoors. Entity "Lighting system" also has its own attributes: "Level of illumination", "Light sources". Entity "Security system" includes system of protection from leakings and ignitions, system of automatic fire extinguishing and video surveillance system. In addition to inheriting from the entity "Life-support system" attributes this entity has its own attribute "State of security", which can take one of two values: "There is a safety violation" or "Violations of safety is not present".

Abstract entity "Sensor" is the parent for entities corresponding to all types of system sensors. It has the following attributes: "Name", "Manufacturer" and "Cost".

Entities "Motion sensor" and "Closing/opening sensor" in addition to inherited attributes have their own attribute "State", which detects movement in the room, closing/opening of the object.

Entity "Leakings sensor" corresponds to a sensor, which is created to detect emergency situations associated with water or gas leakings. This entity has its own attribute "Pressure level".

Entity "Fire sensor" in addition to parent's attributes has its own attribute "Sensitivity", which determines the level of sensor sensitivity to smoke blanketing and temperature changing.

Entity "Temperature and humidity sensor" presents a device, which is responsible for readings of temperature and level of humidity. This entity has its own attributes "Temperature" and "Humidity level".

Entity "System management tool" defines one of the devices, which allow the user to send commands to the central control panel and to inspect the operation of the system. This entity has the following attributes: "Name", "Tool Type" (remote, voice, touch panel), "Manufacturer" and "Cost".

Entity "Central panel" describes the central element of the "Smart House" system, it receives all necessary information from sensors, and it is the center of management of all system components. With a central panel the user interacts through "System management tool". Attributes of this entity are "Manufacturer", "Cost" and "System Components".

After describing of all entities of a metamodel it is necessary to define the relationships between them. The metamodel contains following associations:

- "Send information" is the unidirectional relationship connecting the abstract entity "Sensor" with concrete entity "Central panel".
- "Interact" is the bidirectional relationship describing the interaction of the abstract entity "Life-support system" and concrete entity "Central panel".
- "Fulfill control" is the unidirectional relationship connecting the entities "System management tool" and "Central panel".

In addition to associations the metamodel contains nine inheritance relationships "Is".

Fig. 5 shows one of many possible models of "Smart House" system, constructed in MetaLanguage system with the usage of designed DSML.

ISBN: 978-1-61804-236-1

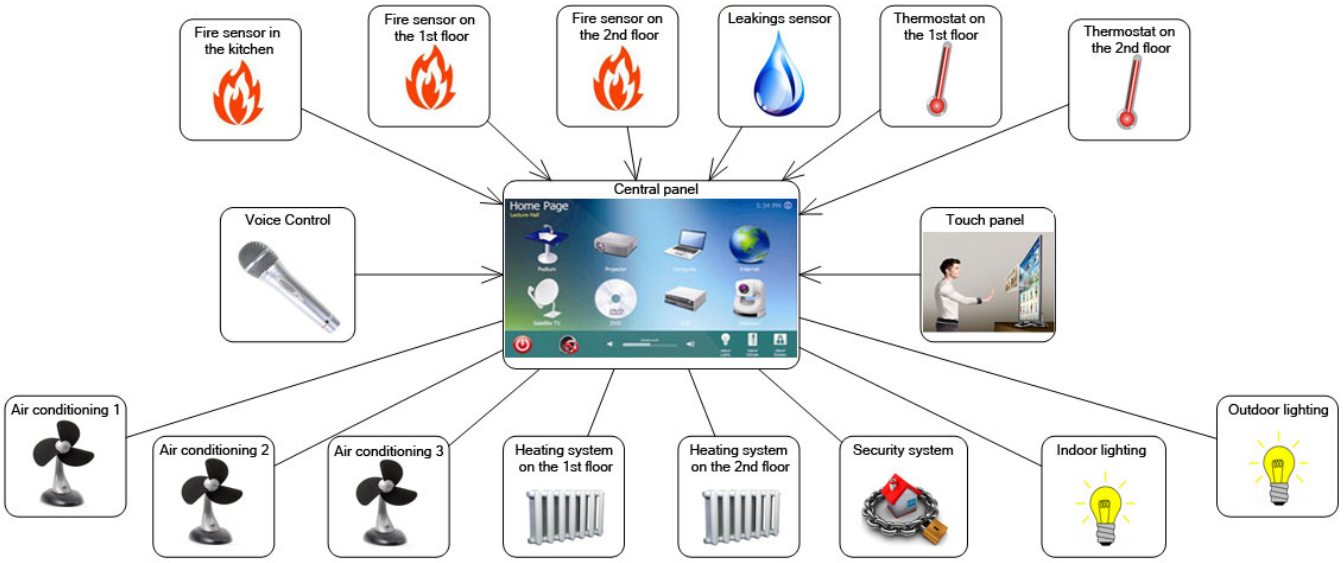


Fig. 5. Model of "Smart House" system

VI. Conclusion

Approaches to development of tools for visual domain-specific modeling languages creation are considered.

The MetaLanguage system allows to describe domain-specific languages, to create models with their usage and to fulfill transformations of the created models in other textual and graphical notations.

This language workbench is simple to use, therefore not only professional programmers, but also domain experts, for example, business analysts, can work with this tools.

For unified models creation the mathematical model – graph grammars based on pseudo-metagraphs – was constructed. This formalism has allowed to describe basic elements of metalanguage and algorithms, which are used at its functioning: algorithms for creation/modification of domain metamodels and models, algorithms for vertical and horizontal models transformations, algorithms for constraints checking.

The MetaLanguage system was approved at the creation of DSMLs and models for several domains (administrative regulations, queuing systems, etc.).

References

[1] 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.

[2] J. Hutchinson, M. Rouncefield, J. Whittle, "Model driven engineering practices in industry," in *Proc. of the 33rd International Conference on Software Engineering*, New York, 2011, pp. 633–642.

[3] W. J. Dzidek, E. Arisholm, L. C. Briand, "A realistic empirical evaluation of the costs and benefits of UML in software maintenance," *IEEE Transactions on Software Engineering*, vol. 34, pp. 407–432, 2008.

[4] M. Velter. (March 2011). MD*/DSL best practices Update March 2011. [Online]. Available: http://www.voelter.de/data/pub/DSLBestPractices-2011Update.pdf.

[5] K. Balasubramanian, A. Gokhale, G. Karsai, J. Sztipanovits, E. Neema, "Developing applications using model-driven design environments," *Computer*, vol. 39, pp. 33–40, 2006.

[6] M. Erwig, E. Walkingshaw, "A DSL for explaining probabilistic reasoning," in *Proc. of the 2nd International Conference on Software Language Engineering*, Berlin, 2009, pp. 164–173.

[7] R. Walter, M. Masuch, "PULP scription: a DSL for mobile HTML5 game applications," in *Proc. of the 11th International Conference on Entertainment Computing*, Berlin, 2012, pp. 504–510.

[8] M. Fowler. (June 2005). Language workbenches: the killer-app for domain specific languages? [Online]. Available: http://www.martinfowler.com/articles/languageWorkbench.html.

[9] S. Kelly, "Comparison of Eclipse EMF/GEF and MetaEdit+ for DSM," in *Proc. of the 19th Annual ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications at OOPSLA 2004*, Portland, 2004, pp. 87–96.

[10] T. Ozgur, "Comparison of Microsoft DSL Tools and Eclipse Modeling Frameworks for domain-specific modeling in the context of the model-driven development," master thesis, Karlskrona, Blekinge Institute of Technology, 2007.

[11] 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 2009*, Orlando, 2009, pp. 147–153.

[12] J.-P. Tolvanen, R. Pohjonen, S. Kelly, "Advanced tooling for domain-specific modeling: MetaEdit+," in *Proc. of the 7th OOPSLA Workshop on Domain-Specific Modeling at OOPSLA 2007*, Montreal, 2007, pp. 48–55.

[13] S. Cook, G. Jones, S. Kent, A. C. Wills, *Domain-specific development with Visual Studio DSL Tools*. Reading. Addison-Wesley, 2007, 560 p.

[14] R. C. Gronback, *Eclipse modeling project: a domain-specific language toolkit*. Reading: Addison-Wesley, 2009, 706 p.

[15] A. N. Terekhov, T. A. Bryksin, YU. V. Litvinov, "QReal: platform of visual domain-specific modeling," *Software engineering*, vol. 6, pp. 11–19, 2013.

[16] A. Basu, R.W. Blanning, "Graphs, hypergraphs, and metagraphs," in *Metagraphs and Their Applications*. New York: Springer US, pp. 1–12, 2007.

[17] B. Courcelle, "Recognizable sets of graphs, hypergraphs and relational structures: a Survey Developments in Language Theory," in *Lecture Notes in Computer Science*, pp. 1–11, 2005.

[18] J. Power, K. Tourlas, "Abstraction in reasoning about higraph-based systems: foundations of software science and computation structures," in *Lecture Notes in Computer Science*, pp. 392-408, 2003.

[19] A. O. Sukhov. (March 2012). Analysis of formalisms for visual modeling languages description. *Modern Problem of Science and Education*. [Online]. Vol. 2. Available: http://www.science-education.ru/102-5655.

Implications of Modern Communication Technologies on Workforce Commitment

Marcus Scholz, Marián Zajko

***Abstract*—**The purpose of this work was to determine the impact of modern communication technologies on the social behaviour of employees. A survey-based descriptive research design was used. The study was carried out among front-end employees of energy supplier companies across Germany. The survey data of the questionnaires were collected from July to September 2013. Stepwise correlation and regression analysis and one sample t-tests were used to confirm the related hypotheses. The findings of the study indicated that the use of Social Media has a high positive correlation with the commitment of the employees. The current investigation contributed to improve the understanding of employee motivation and commitment issues and suggested, a new measurement method of the impact of Social Media use in companies.

***Keywords*—**customer care center, energy supplier, organizational commitment, social media.

I. INTRODUCTION

Ten years ago Social Media (SM) or Web 2.0 started its success in the area of social networks. Never before a mass media technology has reached a greater audience in a shorter period of time. „It took just one year from the time Facebook launched in 2004 to grow to 50 million users“ [1]. Commercial television reached the same amount of households after 13 years and the Internet required three years for it. Today more than 1.2 billion active users around the world have chosen a social network platform for their interactions [2]. This illustrates that SM has become a major part of today's digital world and it is widely accepted among private users. In contrast to the private use of SM which is rather characterized through informal structures [3], the use in businesses is influenced by specific challenges, such as user acceptance, organisational culture, quality management or protection of personal data [4]. In addition to these challenges SM needs to be integrated within the organisational environment in a reasonable way in order to ensure a successful implementation [5]. SM is the first technology that has remarkable *social effects* to the people in the company [3]. Other researchers suppose that the use of SM can probably make a contribution to enhance the employee motivation in a firm [6].

This paper seeks evidence on the impact of SM upon the commitment of the users in the workplaces using SM in their daily work. It will examine the inherent characteristics of this technology in order to help companies to become aware of its potentials. The research scope concentrated on companies in the energy supplier branch restructured due to major legislative changes which opened this market and led to a significant increase in market competition. Therefore both the new and established companies in this branch are keen on seeking new opportunities of winning competitive advantages. The results of this empiric survey in order to analyse the relationship between the use of SM within these companies and the organizational commitment of the employees may suggest promising prospects in this respect.

II. PROBLEM DESCRIPTION

A. State of Research

An increasing number of literature sources examine the general challenges and approaches of technology in the business context. Most of these studies are market research studies, case studies or best-practise papers from institutions and consulting firms [7]–[10]. However, empirical research of this topic has only begun. Some studies from the recent past about the potential of SM use in companies [11]–[13] can be found but they are mostly restricted either to specific functional aspects or they are limited to a company sample. Others investigate just the commitment within an organisation and its interactions within businesses without taking into account specific effects of the SM [14], [15]. The impact of SM on workplaces and its influence on employees' behaviour towards the organization is currently less explored. General presumptions state, that the use of SM will make a contribution to enhance motivation and productivity in the firms [6]. However, there is a lack of empirically verifiable statements about the influence of the implementation of SM within companies on the behaviour and motivation of employees. The published works and available literature mostly examine the fields of application, benefits of use or success factors of the SM use.

B. Aim of Research

This scientific work aims to clarify the overall research question, how the use of modern communication technology like SM in companies may affect the commitment of the workforce. Based on the theoretical considerations about the SM characteristics and its impact on intrinsic motivations a positive influence on the employee commitment may be expected. Moreover a tendency may be assumed that the

ISBN: 978-1-61804-236-1

higher the level of SM use the higher level of organizational commitment may be achieved.

III. Research Methods

A. Approach

In order to get the most reliable data within a given time and cost limit, the data collection has combined a structured personal interview and a survey questionnaire based on a professional online-survey tool.

The initial interviewing of companies management took place as face-to-face or telephone interviews. The second questionnaire on workforce was carried out as an online-questionnaire. Each employee was invited by email to the research survey. After logging-in with a special company access-code the questionnaire was presented on the screen. Reasons for using an online-based method were a better handling of the digital data for analysis with the tools and the possibility to carry out plausibility checks, e.g. prevent from missing values. Furthermore, in a comparison with a conventional paper-based questionnaire the online-survey allows better follow-up activities, such as giving intermediate information and usually higher response rates [16].

B. Research design

The empiric research design consisted of two main parts:

1) The structured interviews of the management of energy supplier companies and
2) The survey-questionnaire among the employees in the customer care centers of these companies.

At first the management of the participating energy suppliers were interviewed to obtain some key statistical data and evaluate the level of SM use in their companies. Then the permission to proceed with the employee questionnaire was obtained and survey scheduled. The management informed the workforce about the purpose and conditions of the survey in advance. The employee survey was focused on the collection of demographical data and commitment values of the workforce.

C. Questionnaires

The aim of the management interview was to figure out the specific level of SM implementation within the respective company. Because of a lack of useable standardized scales, an own questionnaire was generated. The aim of the examination was an objective, reliable and valid scale to measure the use of SM of companies. It was assumed, that there were three essential fields of the SM use:

- Environment in which the SM is used with focus on factors preventing and motivating employees to use SM,
- Level of external SM use focusing on customers and often driven by the sales and marketing departments, and
- Level of internal SM use, focusing on internal communication, collaboration and networking among the workforce.

Integration of these three components generates the Social Media Score (SMS) of a company [17]. The SMS of a company is defined as a value that characterises the overall SM implementation of a firm, considering the level of external and internal uses, as well as the specific environment where the utilization is taking place. In order to be able to measure the SM components the questionnaire consists of 38 items. If there is a low SM usage or activity at all and the environment of its use is missing, a company will achieve few points. With more internal and external use and an appropriate environment, the points achieved will increase. The score of one hundred represents the highest possible level of SM use.

The employee questionnaire was primary designed to measure the employee commitment taking into account the commitment model of Allen and Meyer (1990). Each of the three commitment dimensions reflects one specific commitment type: the Affective Commitment; the Continuance Commitment and the Normative Commitment. For this survey the Affective Commitment scale (Commitment_AC), measuring the emotional side of the respondent, was applied.

The respondents had to choose if they agreed or disagreed with the statements using the seven-point Likert scale from 1– fully agree to 7–fully disagree. In the questionnaire the German translation of the items was applied [18].

IV. DATA COLLECTION

The steps to achieve the objectives of this research work are closely related to the standard empiric research process in the field of economics. The research phase of data collection covers the following steps:

1) Identify and involve participants
2) Realise data collection
3) Evaluate data
4) Process the results.

In the first step of data collection potential companies were identified and invited to participate in the survey.

The data on the participating firms were collected over a 3 months period from July to September 2013. Out of 14 companies participating in this survey, 301 employees took part in the subsequent employee questionnaire survey. After completion of the data collection, the evaluation of the data was proceeded, e.g. screening for consistency, rejecting low-quality data and inverting several items to reach test conformity. After analysing the data by descriptive statistics and taking a deeper look into the research question and the hypotheses, the SMS of the companies was identified and the summary for each firm was prepared and presented in an individual report to each participating company. As a benefit for the survey participation the company received an insight into its individual status of SM implementation, its ranking within the benchmark of all participating firms, as well the recommendations for its further activities in the field of SM use.

ISBN: 978-1-61804-236-1

V. Analysis

A. Correlation analysis

The statistical software package SPSS® (version 16) was used to perform all statistical procedures, e.g. correlation and regression analysis, calculate mean scores, standard derivation and other parameters.

In particular, a correlation analysis was performed to determine the relationship of the average Affective Commitment values of the participating companies from the employee questionnaire to its SMS value as it was determined in the management interviews. Since both variables did not differ significantly from a normal distribution and could be treated like an interval scale, the correlation for this analysis was calculated by the Pearson correlation coefficient [19]. This analysis tests on the relation between two variables and can be viewed as a random or systematic from a statistical perspective. Positive correlation coefficients thereby indicate that higher scores in one variable co-occur with higher scores in the other variable and likewise lower scores with lower scores. Negative correlation coefficients on the other hand indicate that higher scores in one variable co-occur with lower scores in the other variable and vice versa. The assumption predicts a positive direction of the correlation between affective employee commitment and the SMS, so the analysis was conducted with one-tailed testing. Here, the correlation analysis showed a high positive and highly significant correlation of the Affective Commitment to the overall SMS (Pearson r=0.797, p<0.001, N=14). Furthermore the analysis of the SMS subscales was carried out to examine any specific relationship. All three subscales showed a positive correlation to the SMS. The highest correlation was observed with the internal SM usage (Pearson r=0.712, p=0.002, N=14), followed by the Environment (Pearson r=0.523, p<0.028, N=14) and the external SM usage (Pearson r=0.464, p=0.048, N=14) (Table 1).

Table 1 Correlation of Affective Commitment (AC) and SMS

		Subscale External use	Subscale Internal use	Subscale Environ ment	Total Social Media Score
AC	Pearson correlation coefficient r	.464*	.712**	.523*	.797**
	Significance p (1-tailed)	.048	.002	.028	>.001
	N	14	14	14	14

*p < .05 (1-tailed) significant, **p < .01 (2-tailed) significant.

B. Partial correlation analysis

Since it is likely that according to the theory new companies will have a tendency to use new technologies earlier, the effects here are confounded by general differences between older and newer companies. Therefore, the same analysis was conducted with a correction for the influence of the companies' age as the confounding variable, using the partial correlation analysis. Here, the positive correlation between Commitment_AC and the overall SMS (Pearson r=0.566, p=0.03, df=11), remained significant (Table 2).

Table 2 Partial correlation of Affective Commitment (AC) and SMS

		Subscale External use	Subscale Internal use	Subscale Environ ment	Total Social Media Score
AC	Pearson correlation coefficient r	.22	.38	.26	.566*
	Significance p (1-tailed)	.23	.1	.2	.03
	Degree of freedom (df)	11	11	11	11

*p < .05 (1-tailed) significant, **p < .01 (2-tailed) significant.

With respect to the assumption it can be concluded that the degree of the employee's Affective Commitment and a company's SM usage indeed correlate positively and that these effects persist, even if the influence of the companies' age is being controlled for.

C. Regression analysis

Whereas the aim of the correlation analysis was to identify the strength of a connection between two variables, the following regression analysis revealed the kind of this relationship. The scatter plot chart illustrates the data of the two variables: SMS and Commitment_AC. It can be observed that both variables are developing in the same direction and approximately follow a linear trend (Fig. 1).

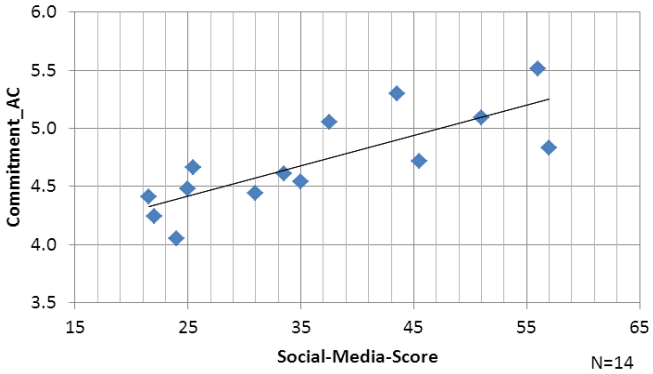


Fig. 1 Graphics of Affective Commitment and SMS

With a regression analysis a consistent linear trend of the values Commitment_AC and the SMS was measured. The SMS significantly predicted Commitment_AC scores (b=0.026, t(14)=4.568, p<0.001) and the value of the Commitment_AC could be calculated by the formula:

$$AC = 0.026 \times SMS + 3.762$$

This linear regression model is able to explain a significant proportion of variance in the Commitment_AC ($R2=0.604$, $F(1, 12)=20.87$, $p<0.001$). Altogether these results allow the conclusion, that the assumption can be confirmed even for the total Social Media Score and especially for the subscale of internal SM use.

VI. Results

The findings of this research clearly indicate that the use of SM positively and highly significantly correlates with the Affective Commitment of the employees. As a measuring instrument the Social-Media-Score was developed to enable an insight into selected aspects of the current status of SM use in the group of energy supplier companies. The investigation proved that the higher the level of SM use in the sample group the higher the Affective Commitment of the employees was.

Consequently, companies that implement the corporate use of SM will have higher personnel benefits in the form of higher dedication, loyalty and motivation of their employees. Some other researchers support these findings. They argue that the development of future information and communication technology that supports people's psychological flourishing by honouring individual and cooperative ideas will enhance the well-being and the quality of life [20]. By providing a work environment where employees can generate and share their knowledge, companies can benefit from an improved information management [21] where superiors have no longer the monopoly of information, that have to be controlled, selected and spread to the employees [22]. In such an environment, the direction of the information flow is no longer only in one-way but rather multi-layered and complex. Not any longer the management automatically knows more than the employees. The resulting participation and involvement of employees additionally change the role of the management from controlling and supervising to moderating and organizing.

References

1] McKinsey Global Institute. *The social economy: Unlocking value and productivity through social technologies*. http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_social_economy, p. 22.

[2] comScore. *It's a Social World: A Global Look at Social Networking.* http://www.comscore.com/Insights/Blog/It_s_a_Social_World_A_Global_Look_at_Social_Networking.

[3] Jahnke, I. In: *Handbook of Research on Socio-Technical Design ans Social Networking Systems*. Whitworth, B., Moor, A. de, Eds.; IGI Global, 2009; Vol. *Chapter L*; pp. 763–778.

[4] Richter, A.; Bullinger, A.C. *Enterprise 2.0 – Present and Future.* [Orig.: Enterprise 2.0 – Gegenwart und Zukunft]. Proceedings of the Multikonferenz Wirtschaftsinformatik (MKWI 2010), p. 741.

[5] Koch, M.; Richter, A. *Enterprise 2.0: Planning, Implementation and successful Use of Social Software in Enterprises.* [Orig.: Enterprise 2.0: Planung, Einführung und erfolgreicher Einsatz von Social Software in Unternehmen]. 2nd ed.; Oldenbourg Wissenschaftsverlag GmbH: s.l., 2009, p. 38.

[6] Komus, A. *Social Software a organisational phenomenon? Applications for businesses.* [Orig.: Social Software als organisatorisches Phänomen? Einsatzmöglichkeiten in Unternehmen]. *HMD Praxis der Wirtschaftsinformatik*, 2006, *252,* 36–44.

[7] Berlecon Research. *Enterpise 2.0 in Germany: Distribution, Chances and Challenges.* [Orig.: Enterpise 2.0 in Deutschland: Verbreitung, Chancen und Herausforderungen], A study of CoreMedia GmbH.

[8] IBM Corporation. *The new collaboration: enabling innovation, changing the workplace.*, 2008.

[9] DETECON. *Customer Care of the Future. With Social Media and Self Services towards new Customer Autonomy.* [Orig.: Kundenservice der Zukunft. Mit Social Media und Self Services zur neuen Autonomie des Kunden]. http://www.detecon.com/de/publikationen/studien/download.html?unique_id=47815.

[10] Ernst & Young. *Social media strategy, policy and governance.* http://www.ey.com/Publication/vwLUAssets/Social_media_strategy_policy_and_governance/$File/Social_media_strategy_policy_governance.pdf.

[11] Benlian, A.; Hilkert, D.; Hess, T. *eCollaboration with Social Software in the global software development.* [Orig.: eCollaboration mit Social Software in der globalen Softwareentwicklung]. *HMD Praxis der Wirtschaftsinformatik*, 2009, *267,* 37–45.

[12] Blaschke, S. In: *Web 2.0*. Blaschke, S., Ed., 1st ed.; Vieweg+Teubner: Wiesbaden, 2009; pp. 183–203.

[13] Räth, P.; Schwaab, J.A.; Smolnik, S.; Urbach, N. *Weblogs used for internal collaboration of the GTZ.* [Orig.: Weblogs in der internen Zusammenarbeit der GTZ]. *HMD Praxis der Wirtschaftsinformatik*, 2009, *267,* 27–36.

[14] Moser, K. *Commitment in Organisations,* 1st ed.; Huber: Bern, 1996.

[15] van Dick, R. *Commitment and Identification with Organisations.* [Orig.: Commitment und Identifikation mit Organisationen]*;* Hogrefe: Göttingen, 2004.

[16] Bungard, W.; Müller, K.; Niethammer, C. *Employee survey – what next?* [Orig.: Mitarbeiterbefragung - was dann …?]. Springer: Heidelberg, 2007, p. 43.

[17] Scholz, M.; Zajko, M. *Social-Media-Score – a tool for measuring the use of Social Media in businesses.* In: Proceedings of the 8th WSEAS International Conference on Business Administration (ICBA '14). WSEAS Press, Ed.: Tenerife, Spain, 2014; pp. 57–60.

[18] Schmidt, K.-H.; Hollmann, S.; Sodenkamp, D. *Psychometric attributes and validity of a German version of the "Commitment" - questionary from Allen und Meyer (1990).* [Orig.: Psychometrische Eigenschaften und Validität einer deutschen Fassung des "Commitment"-Fragebogens von Allen und Meyer (1990)]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 1998(19), 93–106.

[19] Bortz, J.; Schuster, C. *Statistics for Human and Social Scientists.* [Orig.: Human- und Sozialwissenschaftler]. 7th ed.; Springer-Verlag: s.l, 2011, p. 153.

[20] Seligman, M.E.P. *Flourish: A Visionary New Understanding of Happiness and Well-Being*, 2012, p. 94.

[21] Hofmann, J. In: *Enterprise 2.0*. Eberspächer, J., Holtel, S., Eds.; Springer-Verlag Berlin Heidelberg: Berlin, Heidelberg, 2010; pp. 53–61.

[22] Buhse, W.; Newton, T. *Enterprise 2.0,* 3rd ed.; Rhombos-Verl.: Berlin, 2010.

Software Architecture for a System Combining Artificial Intelligence Approaches for Ground Station Scheduling

Michele M. Van Dyne, Costas Tsatsoulis

***Abstract*—** Scheduling of contacts between space vehicles (SVs) and ground stations is of extreme significance since it is essential for data transmission to and from satellites, vehicle maintenance, and orbit tracking and maintenance. We looked at the problem of scheduling contacts between SVs and the U.S. Air Force's Satellite Control Network (SCN). To address the scheduling problem, our work combines case-based reasoning, rule based systems, and generate-and-test techniques, all adopted from artificial intelligence. Our system creates a preliminary, daily SCN schedule with between approximately 500 to 1500 contact requests. The goal is to create a schedule with as few conflicting contact requests as possible, which is then finalized by expert schedule planners. We evaluated our system looking at its performance using only one scheduling algorithm and also using a combination of the algorithms. The system was tested on real SCN schedules and it achieved an average of 75.3% conflict-free over all SCN schedules tested. We also tested the system on schedules created by experts and which contained scheduling conflicts that the experts could not resolve; in these tests our system managed to resolve on average 44.4% of these conflicts, showing performance better than human expert schedulers. This paper addresses the software architecture of our system.

***Keywords*—**Artificial intelligence, case-based reasoning, generate-and-test, rule-based systems, scheduling.

I. Introduction

OUR work looked at the problem of scheduling contacts between space vehicles (SVs) and the U.S. Air Force's Satellite Control Network (SCN). Task scheduling of the SCN is of extreme significance to the Air Force since it is essential for data transmission from and to satellites, vehicle maintenance, and orbit tracking and maintenance. Mission planners plan contacts between their SVs and SCN ground stations.

Complexity arises from the fact that some satellites require equipment or capabilities that are not available at all ground stations. So, when scheduling, one must keep track of the availability of the required support equipment. Additionally, set-up times to configure the equipment must be considered as part of the time required to provide the support. Finally, ground stations themselves require periodic maintenance or emergency repairs.

This work was supported in part by US Air Force Contract # F29601-98-C-0042.

Currently support requirements are expressed and submitted to the scheduling system as Program Action Plans (PAPs). PAPs may be used to specify time windows, support criteria, late starts or early stops, or support preferences such as a required antenna side or unacceptable equipment. PAPs are written in a simplified and ad hoc language.

The challenge of scheduling is to create a schedule that satisfies the needs of the users while not violating any of the constraints inherent in the SCN. A good schedule must achieve as many of the following objectives as possible:

- Optimize network utilization;
- Maximize the number of satisfied requests;
- Satisfy all high-priority requests; and
- Ensure that no SV is denied too many consecutive requests, where "too many" is program dependent.

Human expert schedulers use a number of heuristics to produce good, flexible schedules. Schedules constructed using these principles tend to be easier to modify when real-time changes are required.

In addition to the heuristics, the schedule has to adhere to many constraints and priorities. Constraints may also be flexible and defeasible. For example, high-altitude satellites are more flexible in their scheduling requirements, and turnaround or maintenance down times are padded and can be negotiated down to achieve a workable schedule.

Scheduling is done for different time frames with the shortest one being the 24 hour schedule. This schedule must be conflict-free and is produced manually in a series of steps, described in detail in [1].

The real-time schedule revisions are driven by events that lead to changes in real-time: ground station outages and satellite emergencies. When the support schedule is changed, notification is transmitted primarily by phone calls or face-to-face communications.

SCN Scheduling needs increased automation to deal with the following problems:

- Scheduling SCN assets is a difficult and complex task.
- Priorities are not clearly stated and not uniform from station to station or satellite to satellite.
- The current scheduling process is manpower intensive.
- The input of scheduling data and the manipulation of the schedule is manual, and the process of schedule deconfliction requires significant amounts of effort.

ISBN: 978-1-61804-236-1

To address the problem of deconfliction we developed the ICARUS system (Integration of CAse-based Reasoning and Utility theory for Satellite schedule resolution). The basic technology of ICARUS is case-based reasoning (CBR), i.e. acting intelligently (in this case, performing task deconfliction) using previously successful experiences. CBR is well suited to the deconfliction task since expert schedulers often use the same strategies used in previous deconfliction sessions.

At the same time our research established that there are two more ways by which expert planners deconflict satellite contact schedules: First, there are some well-defined rules that experts use to deconflict schedules. These rules are simple and address simple conflicts, but they are also powerful in that they can address a large number of conflicts.

Second, when experts do not know or cannot design a solution, they revert to "generate-and-test" problem solving. Basically, they try a number of deconfliction techniques hoping that one of them will work. Such problem solving may sound random, but in reality, the techniques used are few, focused, and are developed by decades of experience. These techniques offer an alternate way of solving difficult scheduling problems.

ICARUS does not rely only on case-based reasoning to deconflict schedules. Our experience with domain experts showed that they combine problem solving techniques, and so does ICARUS.

Given a requested schedule that may contain hundreds of conflicts, ICARUS will apply deconfliction rules acquired from experts, will try different changes to the schedule in a generate-and-test mode, and will also use case-based reasoning for deconfliction. ICARUS allows the user to turn on/off the three different deconfliction methodologies (CBR, rules, and generate-and-test), and to perform an analysis of the efficacy of each methodology.

ICARUS was applied to real SCN schedules that had been requested by personnel responsible for particular satellites, and which had hundreds of conflicts. We also applied ICARUS on schedules that had been deconflicted manually by human experts, to see whether an automated system would improve on the performance of experts.

II. RELATED RESEARCH

The area of satellite scheduling can be broken into two major sub-areas. The area addressed in this research is that of space-ground communications, often called the satellite range scheduling problem. The other main area of research in satellite scheduling is that of scheduling tasks on the satellite itself, often called satellite mission planning.

Many different approaches have been investigated in the area of satellite range scheduling. A relaxed version of the satellite range scheduling problem occurs in the area of non-commercial, primarily academic, satellite projects. As described by Schmidt and Schilling [2], under these conditions, ground stations are generally more flexible, contact windows can be shifted to other participating ground stations, time limits are not as strict as those in paid contact situations, and communications are not restricted to a single time window. Schmidt and Schilling describe two approaches to optimize schedules under these conditions: branch and bound, and hill climbing. Their results show that in a test with stations located in four countries, all requests were satisfied after an initial set of requests showed 42 conflicts, and in general, requests were satisfied in an equitable manner. It is not clear, however, how the two approaches were combined, if at all, in producing these results.

Marinelli, et.al. [3] addressed the satellite range scheduling problem using a Lagrangian heuristic. They framed the problem as a multiprocessor task scheduling problem, an approach originating in the operating systems domain. Their approach allowed a relaxation of constraints, which resulted in near optimal performance on large scale test problems.

Yang and Xing [4] combine learnable ant colonies with a knowledge model to improve scheduling performance for the satellite range scheduling problem. The ant colony searches the feasible domain while the knowledge model looks at previous iterations and discovers information that can then be used by subsequent iterations of the ant colony. They tested their approach on 40 generated instances and found that tasks with high priority were consistently scheduled first, while lower priority tasks may have had limits placed on their time windows. Most tasks were scheduled, and schedules produced a high utilization rate and a balanced load.

Howe, et.al. [5] discuss the issues associated with satellite range scheduling and introduce an initial software framework as a basis for approaching the problem. Some of the issues they discuss are that automated scheduling will never be able to completely generate schedules, and that human intervention will always be required, because the problem is over-constrained, and information can become available to human experts that will not be available to an automated system. Furthermore, the nature of the problem is such that an objective optimization function may not be realizable. It is difficult to assign a meaningful weighting and not all the information needed to resolve a conflict is available to an algorithm. Their initial approach involves framing the problem as a job shop scheduling problem and using slack-based and texture-based heuristics coupled with different search algorithms. They use a problem generator to generate realistic, though not real, problems. They also use a web-based interface so that humans using the system at different locations have access to the same system.

Barbulescu, et.al. [6] build on previous work on the satellite range scheduling problem and prove several interesting results. First, they show that the problem is NP-complete. They also show that the results from a reduced problem, that of single resource scheduling, do not generalize into the multi-resource problem. Simple heuristic approaches perform well on "easier" problems, circa 1992, but as the number of requests has grown larger, these approaches do not scale up. Finally, they show that a genetic algorithm approach yields the best results on

larger, more complex problems, which are more representative of present-day communications traffic.

As recently as the early 2000's, a system called ASTRO was in use as the satellite range scheduling system for SCN resources. ASTRO was a set of tools for compiling, storing, displaying, and manipulating SCN resource requests and the resulting schedules. ASTRO was a DOS-based system that allowed the human scheduler to enter schedule requests and manipulate this data to produce a network schedule, though it did not automate the decision process. ASTRO featured a large-screen monitor to display the schedule and a sonic pen used to manipulate the schedule [7].

Case based reasoning has not been investigated much in the area of satellite range scheduling, but it has been used in other planning and scheduling areas. Related to ICARUS, described in this research, are case based planning systems, such as CaPER [8], a CBR system that uses high performance computing techniques for fast retrieval. CaPER also attempts to merge plans and to resolve harmful interactions between them. ForMAT uses CBR to retrieve old plans, represents temporal relationships, and assists the user for revisions and re-planning [9]. CABINS used CBR to schedule job shop activities. CABINS represented cases based on the temporal constraints they satisfied and the goals they achieved, and had a constraint-based scheduling component to iteratively repair schedules retrieved by CBR [10].

This paper focuses on the software architecture of ICARUS. A more complete description of the decision methodology contained in the software can be found in [11].

III. ICARUS OPERATION AND ARCHITECTURE

A. Overview

ICARUS takes as input a conflicted schedule and produces a schedule that has been deconflicted as much as possible. One constraint on the schedule input was that our system had to read and parse the schedule format produced by the tool being used, ASTRO. ICARUS was then required to output the deconflicted schedule in the same format. The input/output language is not effective for making scheduling decisions, however, so part of the process of parsing the input files was to generate data in a normalized database format.

Once a conflicted schedule and any necessary additional information was loaded into ICARUS, it iterates over its three deconfliction engines: rule-based, case-based, and generate-and-test. The user can control the number of iterations and which deconfliction engines can be used. The only constraint is that the user cannot change the sequence in which the engines are applied to the schedule.

ICARUS allows the user to view the original and the deconflicted schedules. The system also generates appropriate schedule statistics, such as total number of tasks, conflicts, and visibilities, and complexity of conflicts (i.e. the number of tasks conflicting with another task.)

Each of these processes is discussed below, and the overall architecture is shown in Figure 1.

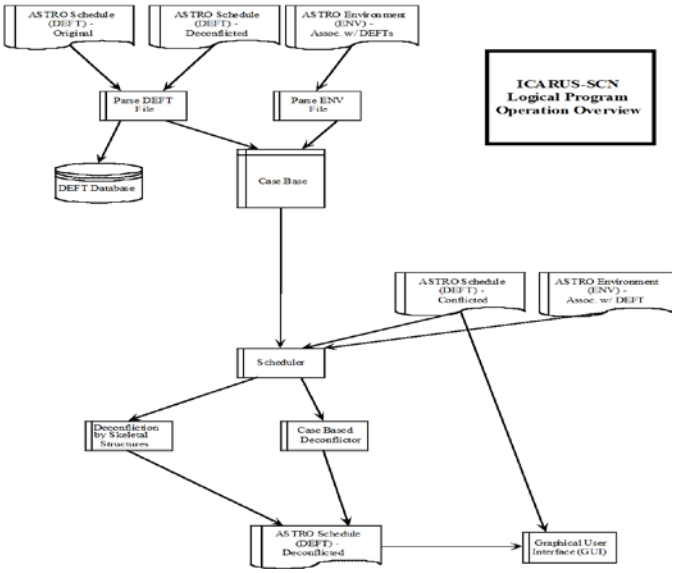


Figure 1: ICARUS Architecture Overview

B. Input/Output – Parser and Inverse Parser

The scheduling assistant tool used by the Air Force SCN was called ASTRO. The underlying language associated with ASTRO was ad hoc, and not structured in a way that it could be operated on programmatically. The first step, then, in building the system was to parse this language into meaningful and structured data. An example of two entries in the DEFT format used by ASTRO are:

```
QK350LION-BSTRNGTCS 03122015000030030800000060000 N N
N          008008          12202C008 STA TRNG, 30 MIN
BLK,    W=1800-2100Z,    NOT    W/OTHER    ANT    D/T..
S008                     ARTS                    AR33
L                1800-2100                 :30/:30/:30
QH905HULA-BSPMI HTS 03121630000400030819000600000 N N
N          013013 8       12166C013 PROTECTED PMI, -
00/+96  HRS  OF  MON/1700Z,  PREF  NO  SHIFT  CROSS  OF
17,01,09Z.           S013           ARTS           AR46
L                                                4/4/:30
N NT     CROSS OVER OK..KR
```

The DEFT files describe scheduling requests, while ENV files describe ground stations and available equipment and resources on those stations, covering the time period for the requests in the associated DEFT files. Both of these files are parsed into an object format which represents a normalized structure for SQL database storage. As an example, task requests have the structure shown in Figure 2.

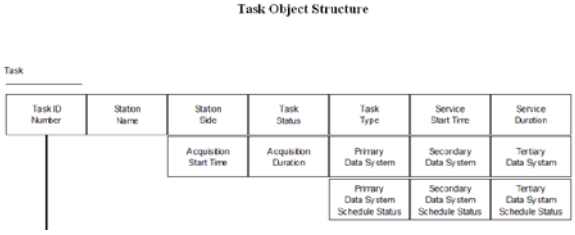


Figure 2: Parsed Task Structure

ISBN: 978-1-61804-236-1

After completion of its deconfliction process, ICARUS performs the inverse parsing of its internal data representation structures to produce a file readable by the ASTRO program.

In addition to file I/O, ICARUS allows user control of certain operational parameters. A user can specify which of the three deconfliction engines, in combination or alone, can be used during the session and how many iterations the deconfliction process is allowed to make before stopping.

C. Database

ICARUS uses the parsed tasks and environments to create a relational database. The database management system used is MySQL. ICARUS can create the database structure if needed, and populates the database with the task requests and environment data of the parsed DEFT and ENV files.

Figure 3 shows the overall layout of the database structure used by ICARUS in performing its deconfliction. While individual relation and field names in the image are not readable, the overall structure of the relations is evident in the diagram.

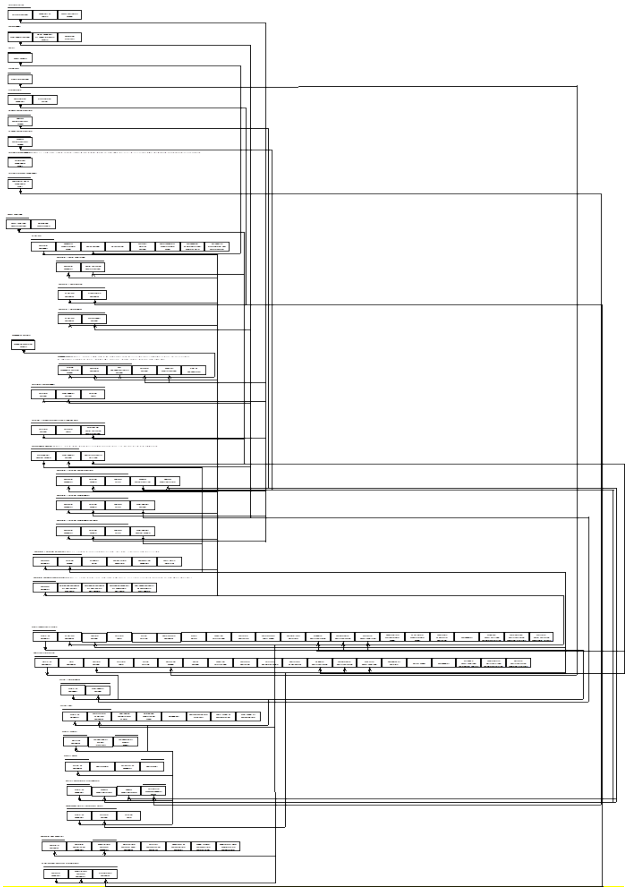


Figure 3: ICARUS Database Structure

D. Deconfliction Engines

1) Case-Based Reasoner

In ICARUS a case contains a description of a conflicted task and the knowledge inside the case stores specific information about how the conflict was resolved. A case base is created by using two schedules for the same set of tasks. The first schedule (the "before" schedule) is the not yet deconflicted set of requests by the Satellite Operations Centers. The second schedule (the "after" schedule) is the deconflicted set of the same set of requests. Cases used in the ICARUS case base are those where deconfliction was performed by expert schedulers. ICARUS identifies the same task in the before and after schedules, making sure that the task had conflicts in the before schedule and has no remaining conflicts in the after schedule.

The case description is a description of the conflicted task. It is a descriptor of the context and specific details of the task, such as equipment it requires, its duration, its time constraints, and so on. It is meant to help the case-based deconflictor identify similar, conflicted tasks. The case description contains the information used in matching tasks, which are actual contacts between a satellite and a ground station.

The solution part of the case consists of the ways in which the conflict was resolved. This information is extracted by studying the task in the before and after schedules and identifying how the conflicted task was changed in the after schedule. The ICARUS case-based reasoner identifies the following changes to a task: 1.) change station; 2.) change station side; 3.) change start time; 4.) change turn-around-time; 5.) change duration; and 6.) change data system.

Given a case base, ICARUS uses it to resolve conflicts. To do so it selects the best case from memory by a weighted matching of all features of a conflicted task against all cases, followed by ranking of the cases by the matching weights. Each feature in ICARUS is given a weight between 0 (not used in matching) and 1. Features that are symbolic and single valued are matched one-to-one (binary match). Features that are numeric and single valued are matched by the value weighted by the inverse of the difference between the two values. Multivalued features are matched as the intersection of matching values (such multivalued attributes are, for example, the equipment list or the list of preferred stations). The matching value is weighted by the feature weight and all weighted values are summed to generate the final matching value for a case.

After the best case is selected, ICARUS attempts to apply the deconfliction solutions found in the case subject to the constraints defined in the task. If a case has more than one potential deconfliction action, ICARUS attempts to perform each one of the actions until it either deconflicts the task, or all steps fail.

Figure 4 shows the overall architecture of the case-based reasoning portion of the ICARUS system.

2) Rule-Based Deconfliction

Rule-based deconfliction is based on the way that expert schedulers initiate deconfliction of schedules. Schedulers attempt to first "slide" a task on the same station side, trying to find an opening wide enough to accommodate the task, and where equipment is available and the task is within its visibility window.

ISBN: 978-1-61804-236-1

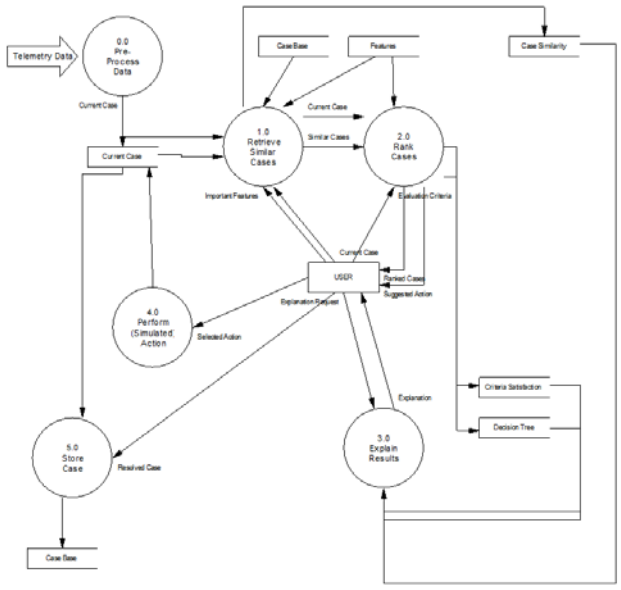


Figure 4: Case Based Reasoner Architecture Overview

ICARUS does the same thing. Without changing anything other than the start time of a task, it slides a task obeying the time constraints and making sure the task is visible and the necessary equipment is available. If the start time is fixed and the task start time has been defined incorrectly, the rule-based deconflictor will move the task to the correct time. The rule-based deconflictor also makes sure the open time it finds will accommodate the defined turn-around-time.

This process is described more formally using set theory constructs. Given are a set of antenna sites $\{S_i\}$, and a set of space vehicles $\{V_j\}$. In general, a vehicle V_j is visible to a site S_i for various time intervals during the day. There are a set of assigned tasks for the sites. This implementation of the rule-based portion of ICARUS attempts to move the time intervals to reduce the number of conflicts. First it reads the tasks for a given site, and for each task, stores the time interval, identification number, and task type. Let this set of time intervals be

$$T = \{t_1, t_2, \ldots, t_n\} \quad (1)$$

These intervals are sorted according to starting time, then any set of sequential overlapping intervals are combined into a single interval. Then the resulting set

$$U = \{s_1, s_2, \ldots, s_n\} \quad (2)$$

is unavailable time space. Initially no task could be moved to this space to eliminate a conflict. The available space is the complement

$$A = U^C. \quad (3)$$

For each task P to be shifted to the available space A, there is an associated visibility set, Vp, which is a set of time intervals, so the available space is refined as

$$A_p = A \cap V_p \quad (4)$$

Calling the i^{th} overlapping set O_i, for each interval i in O_i, we examine the available space A_p, and find the translation seconds to shift I to the closest available space. If this is successful, we subtract the translated interval I_T from A to get a new available space. We continue this until we reach the last element of the overlapping set O_i and can either report success or failure.

3) Generate-And-Test Deconfliction

We noticed that there was a finite set of actions schedulers (and ICARUS) can take to deconflict a task: change station, change side, change service start time, change turn-around-time, and change task duration. Consequently, we added one more deconfliction engine to ICARUS, one that cycles through all possible deconfliction actions until it finds one that resolves the conflict. In other words, this engine generates a possible deconfliction action and tests it to see if it will work. This is the "generate-and-test" deconfliction engine.

The major difference between case-based and generate-and-test deconfliction is that the former elects to perform only the best deconfliction actions based on its experience, while the latter will try all steps.

Each change is tested against the same constraints as the changes performed by the case-based deconfliction engine. The sequence of changes attempted in this approach was established in collaboration with experts, and represents an increasing disturbance of the task. So, ICARUS will first attempt the least intrusive changes, the ones that leave the scheduling requests as unchanged as possible, and will increasingly disturb these requests (within constraints) until a solution is found.

E. Viewer

As deconfliction progresses, ICARUS displays messages and progress bars on the screen to keep the user informed. The user may also view the schedule before and after deconfliction using our schedule viewer. An example of this is shown in Figure 5.

The x-axis is time, here spanning four and a half days. The span depends on the input schedule. The y-axis consists of the satellite stations, read in from the ENV file. The color bars represent the satellite contact tasks (a different color for a different contact), red bars represent conflicts in the schedule, and green bars represent station maintenance tasks.

ISBN: 978-1-61804-236-1

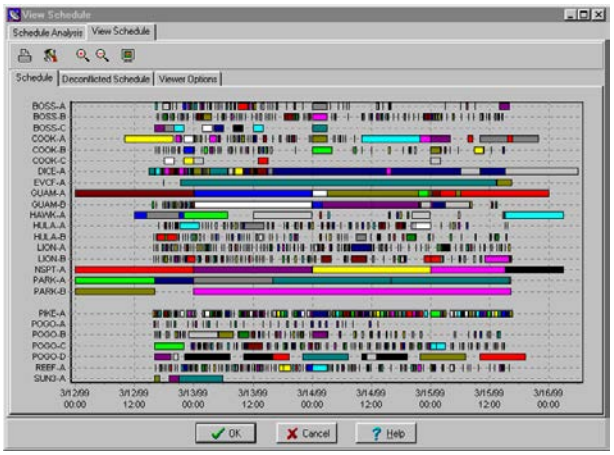


Figure 5: Graphical View of Conflicted Schedule

IV. RESULTS

We evaluated our system on actual SCN schedules and some of the results are shown on Table 1. The schedules started with a number of tasks and conflicts due to conflicting contact requests. ICARUS used all its deconfliction methods sequentially (rule based, CBR, and generate-and-test), and the results of each deconfliction step are listed in the table. For example, the first schedule shown in Table 1 started with 379 conflicts and after applying the rule based deconflictor there were 301 remaining conflicts, after applying CBR there were 300 conflicts, which were then lowered to 244 by generate-and-test. The system iterated three times, and stopped after the conflicts did not change after an iteration cycle.

We tested our system on SCN provided schedules, and after ICARUS the average schedule was 75.3% clear of conflicts. We also tested ICARUS on schedules created by experts and which contained scheduling conflicts that the experts could not resolve; in these tests our system managed to resolve on average 44.4% of these conflicts, showing performance better than human expert schedulers.

Table 1: Sample Deconfliction Results

Tasks	Initial Conflicts	Iterations (Orig→RuleBased→CBR→G&T)	Final Conflicts
562	379	379→301→300→244 244→241→241→239 239→239→239→239	239
717	69	69→63→63→23 23→22→22→22 22→22→22→22	22
535	241	241→138→135→116 116→116→116→116	116
587	300	300→202→199→163 163→161→161→161 161→161→161→161	161
1505	617	617→458→447→352 352→350→350→348 348→348→348→348	348

V. CONCLUSIONS AND FUTURE WORK

Our work addressed an important operational need of satellite control networks: how to resolve conflicting requests for access to the ground station by space vehicles. This problem is different from traditional scheduling or planning ones, since it starts with an existing schedule which corresponds to scheduling requests. These requests are often conflicting, and require correction.

There are two potential extensions to ICARUS, both improving its deconfliction performance. ICARUS could implement hand-offs between ground stations. In other words, a contact task could be shared between two stations, if it could not be fully satisfied at one station. A large number of conflicts can be resolved if a long contact request can be broken into smaller contacts that are distributed over a set of stations. Also, certain resources can be shared by multiple tasks. Allowing sharing of resources and equipment will improve deconfliction performance.

There will almost always be conflicts in every schedule that cannot be resolved because of hard constraints. These scheduling requests are denied by human users, something our system is not allowed to do. Consequently, regardless of improvements to our system, it will never generate a 100% conflict-free schedule.

REFERENCES

[1] Loral Federal Services Corp. 1995. CCSU Resources Scheduling Study Report Contract F04701-91-C-108, CDRL A115.

[2] M. Schmidt, and K. Schilling. 2009."A Scheduling System with redundant scheduling capabilities." International Workshop for Planning and Scheduling in Space, Pasadena, USA. 2009

[3] F. Marinelli, S. Nocella, F. Rossi, & S. Smriglio. 2011. "A Lagrangian heuristic for satellite range scheduling with resource constraints", Computers and Operations Research, 38 (2011), 1572-1583.

[4] K. Yang and L. Xing. 2012. "The Learnable Ant Colony Optimization to Satellite Ground Station System Scheduling Problems", Electrical Review, R. 88, NR 9b/2012, 62-65.

[5] A.E. Howe, L.D. Whitley, L. Barbulescu, J.P. Watson. 2000. "Mixed Initiative Scheduling for the Air Force Satellite Control Network", Second International NASA Workshop on Planning and Scheduling for Space, March 2000.

[6] L. Barbulescu, J.P. Watson, L.D. Whitley, A.E. Howe. 2004. "Scheduling Space-Ground Communications for the Air Force Satellite Control Network", Journal of Scheduling, Vol. 7, Issue 1, pp. 7-34, January.

[7] Loral Federal Services Corp. 1995. Automated Scheduling Tools for Range Operations (ASTRO), Contract F04701-91-C-108, CDRL A058.

[8] Hendler et al. 1994. "Massively Parallel Support for Case-Based Planning," Proc. of APA/Rome Lab Planning Initiative Workshop, Morgan Kaufmann.

[9] Mulvehill, A. 1995. "Reusing Force Deployment Plans," AAAI Fall Symposium on Adaptation of Knowledge for Reuse.

[10] Miyashita, K. and K. Sycara. 1994. "Adaptive Case-Based Control of Schedule Revision," in: Intelligent Scheduling, M. Zweben and M.S. Fox (Eds.), San Francisco: Morgan Kaufmann, 291-308.

[11] Tsatsoulis, C., and Van Dyne, M. " Integrating artificial intelligence techniques to generate ground station schedules", Proceedings of the 2014 IEEE Aerospace Conference, March 1-8, 2014, Big Sky, MT.

Two stage strategy of job scheduling in grid environment based on the dynamic programming method

Volodymyr V. Kazymyr, Olga A. Prila

***Abstract*—**The problem of the effective usage of the grid environment for solving different types of computing tasks of large dimension is researched in the paper. We study the problem of optimal task scheduling at an affordable set of resources on the one hand and the equitable distribution of resources between the tasks that come into the input queue of a centralized workflow management system, on the other hand. Two stage strategy of task scheduling in grid environment that takes into account user-defined QoS requirements, structural features and execution dynamicity of the task is presented. The dynamic programming method application to the workflow scheduling problem is proposed in the paper and the effectiveness evaluation experimental results of the proposed decision are given.

***Keywords*—**Grid, workflow, scheduling, QoS

I. Introduction

Currently grid technologies are actively developed and applied to solving of complex high-dimensional problems. The problem of task adoption for effective execution in the grid environment is rather complex itself. Abstract features of using of the grid-infrastructure for different types of tasks' execution based on their structural but not functional features can be determined.

The actively developing field in the grid-computing is the technology of workflow execution in grid-environment. The workflow represents the task as a sequence of subtasks with certain synchronization scheme. The presence of several parallel blocks in such tasks allows to execute them on different resources for more efficient problem solution. To provide such a solution several factors should be taken into account and the most important of them is the expenses for data exchange between utilized resources. Workflow scheduling is an NP-complete problem in general [1].

Another component that is very important for grid end-users is the ability of a grid system to provide its consumers with the required quality of service (QoS). It results in the need to allocate task on a set of resources which is most suitable for its execution depending on the QoS information provided by the resources. While for non-commercial community grids it is limited to estimate completion time (ECT), commercial utility grids can also operate with costs of calculations and some other parameters.

The paper is dedicated to the research of the aspects of scheduling different types of tasks in grid-environment paying much attention for the workflow scheduling problem. Another research problem is the strategy of the central queue of grid tasks processing according to their priorities and QoS requirements.

The complex two stage strategy for tasks' queue processing and task scheduling next is proposed in the paper.

II. The Formalization Of The Task Of Scheduling Different Types Of Jobs In Grid Environment

One of the factors influencing the performance of the grid-network is planning efficiency. Taking into account the heterogeneity of grid-resources, as well as the structural features of tasks, the following factors should be understood under the scheduling efficiency.

1) Equable load of all the grid computing elements.
2) The minimal tasks' downtime in the run queue.
3) The minimal execution time of tasks on a dedicated set of resources, including the time required for data transfer between computing blocks.

The classification of the tasks which are calculated in the grid-environment according to the structural criterion has been suggested by the authors [2].

The execution effectiveness of the task represented by a single computing unit or a set of consistent, depends on the effectiveness of its program implementation, planning strategies of the low-level grid brokers and local scheduler.

If the task is represented by a set of similar tasks with different input data, scheduling optimization reduces to decomposition of the task according to the current options of grid-infrastructure.

The presence of parallel blocks in workflow tasks allows executing them on different resources for more efficient problem solution. To provide such a solution the expenses for data exchange between utilized resources should be taken into account.

V. V. Kazymyr, Dr. Sc. Prof. Chernihiv State Technological University, Shevchenko street, 95, Chernihiv-27, Ukraine, 14027; e-mail: v.vkazymyr@gmail.com.

O. A. Prila, postgraduate, the assistant lecturer, Chernihiv State Technological University, Shevchenko street, 95, Chernihiv-27, Ukraine, 14027; e-mail: olga.prila1986@gmail.com.

ISBN: 978-1-61804-236-1

The expenses for data exchange can be eliminated by clustering several blocks of workflow for the same resource. There is a concept of linear and nonlinear clustering [1], when serial or parallel blocks are grouped respectively. The optimization problem is reduced to finding the optimal solution between parallelization and clustering.

Workflow task is generally modeled by a precedence-constrained task graph, which is a directed acyclic graph with nodes representing the subtasks and the directed edges representing the execution dependencies between them as well as the amount of communication (see Fig. 1).

For effective workflow scheduling the following subtask parameters must be defined:

{ECT, Memory, {T}},

where ECT - estimated completion time;

Memory – memory requirements;

{T} – set of links to other modules (one-way communication between nodes).

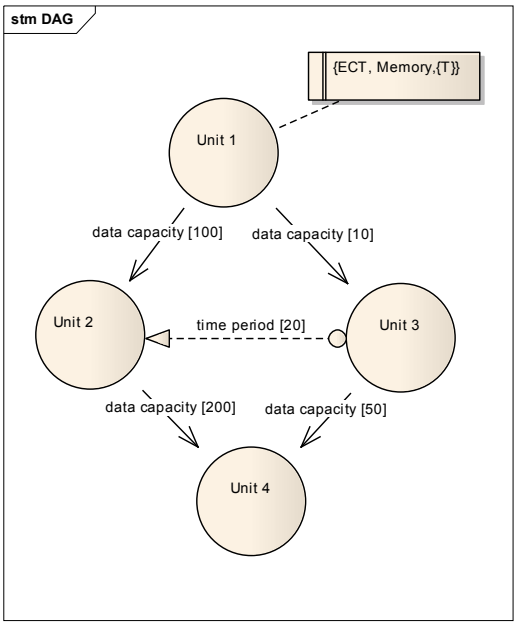


Fig. 1 the example of workflow task structure

Each relationship is defined by the data capacity parameter – the amount of data transferred between the units. The relationship between Unit2 and Unit3 determines the need for periodic synchronization of data between the units that are executed in parallel.

The complex type of the task which includes the characteristics of several types should be considered separately. There may also occur a particular case of type three when the task is decomposed into totally independent parallel subtasks. In the latter case the process of adaptation to grid is significantly simplified due to the absence of necessity of synchronization between the calculation blocks.

For effective execution in the grid-environment the following limitations are imposed for the workflow structure.

1) Lack of loops and branches. These limits are determined by the task structure presentation in the form of an acyclic directed graph. However, the task structure having loops can be converted to DAG by adding an additional level. Branching can be handled at the level of metascheduler through the dynamic approach to planning.
2) High level of task granularity. The dimension of calculations should be much higher in relation to the dimension of the transmitted data [1]. In [3] the granularity problem is defined as follows:

$$g = \min_{x=1:\nu}\{\tau_x / \max_j\{c_{x,j}\}\}, \quad (1)$$

where τ_x – computational complexity of node n_x;

$c_{x,j}$ – dimensionality of the data being transferred between nodes n_x and n_j;

ν – the number of computing nodes of the task.

Grid-network structure can be presented as a complete directed graph where vertices define the resources, and the weight of arcs define bandwidth computer network (see Fig. 2).

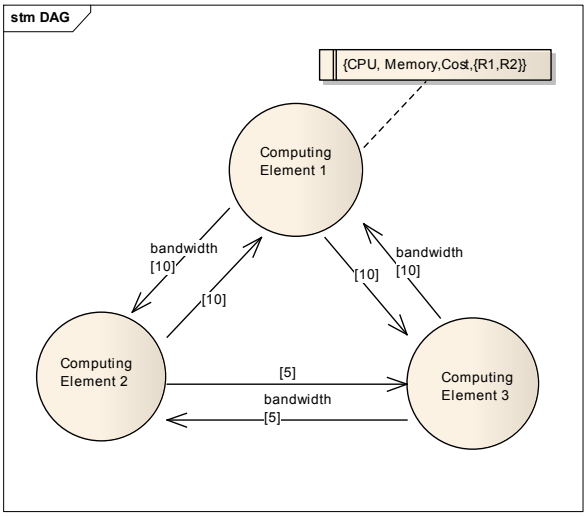


Fig. 2 grid-structure example

Each unit of the grid-net structure is characterized by the following compulsory parameters.

{CPU, Memory, Cost, {R1, R2}},

where CPU – computational power, Memory – memory characteristics, Cost – usage cost, R1 – receive data network bandwidth и R2 – data transmission network bandwidth.

The optimization task presupposes working out an optimal variant of the stream of jobs arrangement on the available set of resources.

The classification of jobs scheduling according to the following criteria is presented in [4]:

$$\{\alpha \,|\, \beta \,|\, \gamma \,|\, \delta\}, \quad \beta = \{\beta_1 \,|\, \beta_2 \,|\, \beta_3\}, \quad (2)$$

where α – determines the characteristics of the distributed environment (homogeneous / heterogeneous);

β – the job specifications and the presence of limitations in the job structure;

β_1 – presence of relation between the job computing units;

β_2 – homogeneity / heterogeneity of the job computing units;

β_3 – presence of time limitations of the job computing unit irrespective of the results of the units connected with it;

γ – determines the optimization criterion and the type of the objective function;

δ – determines the expenditure function of the distributed environment resources interaction in job performance.

According to the classification suggested the job scheduling in the grid-environment is determined the following way

$$\{R \mid PREC, \varnothing, r_i \mid C_{\max} \mid JP\}, \quad (3)$$

where R – determines the heterogeneity of the distributed environment resources; the time of the job computing unit performance is the power function of the distributed environment unit;

$PREC$ – the presence of relations between the job computing units;

$\varnothing$ – the job units have different computing complexity;

r_i – there are time limitations as for the beginning of the job computing unit;

$C_{\max}$ – the objective of the job scheduling is the minimization of the time of job performance;

JP – the expenditures on the distributed environment units interaction are determined by the parameters of the network bandwidth, as well the job specifications (the level of transmission data between the computing units).

The suggested classification does not take into account the multicriterion characteristic of the objective function. However, besides the job performance time, the competitive criterion, computing cost, is significant for the commercial grid-environment. The task of jobs streams scheduling in the grid-environment is generally NP-total task.

At the middleware level grid does not provide full support for the tasks of different types. For instance, ARC Nordugrid (http://www.nordugrid.org/) and gLite (http://glite.cern.ch/), which are on the list of the main providers of grid middleware in EMI (http://www.eu-emi.eu/) and which are widely used by the Ukrainian national grid-infrastructure make use of the following formats of the grid-tasks specifications: JSDL [5], xRSL [6] and JDL [7]. Among the mentioned JDL-format is the only to introduce the notion of the task type (Job, DAG и Collection), which still has certain limitations – the determination of the periodic synchronization between the units and the data transmission levels is impossible. JSDL and xRSL formats provide just means of determining the parameters of some tasks; however, the jobs stream life-cycle, as well as the relations between certain tasks, is not supported.

Middleware grid brokers realize simplified scheduling strategies and do not allow performing jobs stream scheduling taking into consideration the tasks specifications and QoS parameters. For instance, ARC Nordugrid broker realizes the following strategies of the available computing resources selection: RandomBroker, BenchmarkBroker, FastestQueueBroker, DataBroker [6]. The latter means that the mechanisms of complex grid-tasks scheduling are to be realized and arranged beyond the middleware grid level.

III. The Structure Of The Metascheduler Of The Centralized Workflow Management System

An important component of the use of the grid-environment is to provide its consumers with the required level of quality of service (QoS).

In addition to finding the optimal schedule of workflow task on the set of available computing resources, the important aspects of the metascheduler are: a) the strategy of processing the input queue of tasks in accordance with their priorities; b) the choice of the scheduling algorithm according to the structural features of the problem; c) dynamic control of task execution; d) accounting the dynamicity of grid-network as well as the level of quality of service of grid resources.

Below we consider the metascheduler implementation aspects which are an integral part of the centralized workflow management system. Lack of the centralized approach, which consists in the possible occurrence of bottleneck, is assumed to be eliminated through the scalable workflow management system architecture.

This paper introduces a two-stage strategy for task scheduling in grid environment. The first stage involves the processing of the input queue of tasks in accordance with their priorities and QoS requirements. The second phase involves task scheduling at the affordable set of resources taking into account the structural features of the task.

Below we consider the approaches to scheduling the workflow task on the set of available heterogeneous grid-resources as well as strategies for handling the input queue of tasks of different types accounting the dynamicity of their execution.

IV. Dynamic Programming Method Application TO The Problem Of Workflow Scheduling

In [1, 8, 9] the classification and the results of the algorithms' effectiveness and complexity evaluation are given.

The methods of search in the space of states and methods of mathematical programming can produce optimal solutions, but in general are characterized by high computational complexity of the algorithm. Heuristic approaches can give effective solutions in polynomial time, but in general these approaches do not provide the optimal solution, as the average, the worst and the best performance of these algorithms is unknown [1].

Clustering (DSC, CASS-II [10]) and replication (TDS, TANH [11]) approaches aimed at reducing the time required

for data transfer between nodes by placing tasks that require the exchange of large amounts of data on a single resource or duplicate blocks, respectively [9]. The disadvantages of these approaches is the difficulty of accounting the heterogeneity of the subtasks, and the lack of opportunities to use several resources grid-network for parallel blocks task.

An important aspect of the use of commercial grid-environment is the need to optimize the characteristics of mutually exclusive - resource cost and execution time taking into account the significance of the coefficients of each of the characteristics. Most heuristic scheduling algorithms focus on improving one of the criteria. Today, the only workflow scheduling algorithm that solves the multiobjective optimization problem is the LOSS / GAIN algorithm [8].

Many of the existing scheduling algorithms impose some restrictions on the structure of the problem, the structure of grid-network optimization criteria.

Dynamic programming method is one of the methods of mathematical programming, applied to the problem with optimal substructure. Optimal substructure problem assumes that the optimal solution of its constituent smaller subtasks can be used to solve the original problem [12].

The algorithm introduces the concept of levels in the structure of the problem of the "work flow", which is determined by a variety of tasks that can be performed simultaneously at a certain stage of the task. For example, a workflow structure shown in Fig. 1 may be allocated to the following levels: 1) {Unit 1}; 2) {Unit 2, Unit 3}; 3) {Unit 4}.

Optimal solution contains optimal solutions at every level, and, therefore, the task has the property of optimality [16].

The objective function of the algorithm can be determined by several parameters that have some weight. Accordingly, the objective function might look as follows:

$$y = k_1 \cdot time + k_2 \cdot \cos t \quad (4)$$

where k_1, k_2 - user-defined coefficients of QoS parameters;

time – task execution time;

cost - the cost of computing resources usage.

The flowchart of the algorithm is shown in Fig. 3.

As it can be seen from the flowchart, at each level for each allocation variant the optimal solution is saved regarding the allocation cost as well as the cost of interaction with the blocks of the previous levels. Inefficient solutions for each location are discarded and will not be further considered. At the last step of the algorithm the global optimal solution is determined moving backward from the bottom up through the levels. It is recommended to store a copy of the accommodation plan ordered by the value of the objective function, which can be used for dynamic rescheduling problem if necessary.

In the case of existence of the “through the level” communication link the “dummy" block of zero computational complexity is assumed to be added.

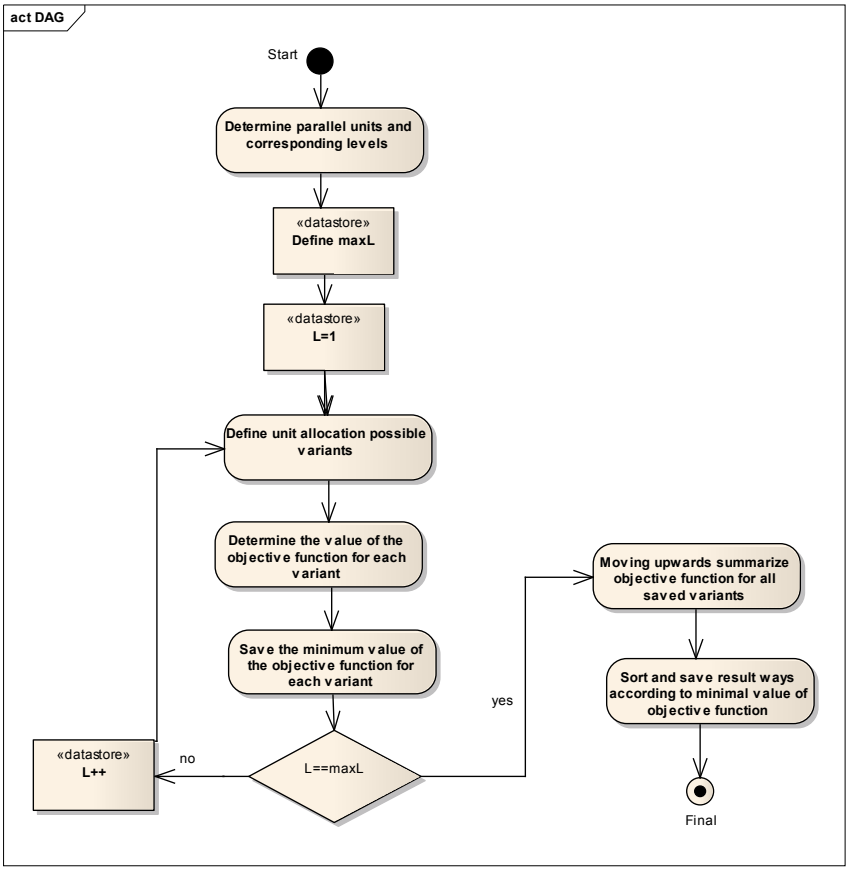


Fig. 3 the QoS-based scheduling algorithm flowchart

Consider the example of the algorithm for the workflow

shown in Figure 2 and the network structure shown in Figure 3.

For the simplicity we take the objective function as

$$f(\text{time}) \rightarrow \min \qquad (5)$$

The input data is represented in a table structure (see Table 1-2).

Table 1. Workflow structure

	1	2	3	4	ECT
1	#	100	10	#	500k
2	#	#	#	200	2000k
3	#	#	#	50	100k
4	#	#	#	#	100k

Table 2. Grid network structure

	0	1	2	CPU
0	0	10	10	100k/time
1	10	0	5	50k/time
2	10	5	0	50k/time

Step 1. Determined levels:

Level 1: Unit 1

Level 2: Unit 2, Unit 3

Level 3: Unit 4.

Step 2. For set №1 define all possible variants of unit allocation:

U1CE1, U1CE2, U1CE3.

Step 3. Determine the value of the objective function for each variant.

U1CE1: 500k / 100k = 5;

U1CE2: 500k / 50k = 10;

U1CE3: 500k / 50k = 10;

Step 4. Save the value of the objective function for each variant.

Step 5. Turn to set № 2. Repeat steps 2-4. First, we define the computational cost for U2 and U3 (for ease of computation), and then for all possible combinations of the location of the previous set of blocks.

a) U2CE1U3CE2: 2000k/100k + 100k/50k = 22;

b) U2CE1U3CE3: 2000k/100k + 100k/50k = 22;

c) U2CE2U3CE1: 2000k/50k + 100k/100k = 41;

d) U2CE2U3CE3: 2000k/50k + 100k/50k = 42;

e) U2CE3U3CE1: 2000k/50k + 100k/100k = 41;

f) U2CE3U3CE2: 2000k/50k + 100k/50k = 42;

The calculated objective function for each combination of set №2 is presented at table 3.

Table 3. The objective function value for each combination of set №2

	U1CE1 (5)	U1CE2 (10)	U1CE3 (10)	Min
U2CE1 U3CE2 (22)	22+5+0+ 10/10	22+10+100/ 10+0	22+10+1 00/10+10 /5	28(U1C E1)
U2CE1 U3CE3 (22)	22+5+0+ 10/10	22+10+100/ 10+10/5	22+10+1 00/10+0	28(U1C E1)
U2CE2 U3CE1 (41)	41+5+10 0/10+0	41+10+0+10 /10	41+10+1 00/5+10/ 10	52(U1C E2)
U2CE2 U3CE3 (42)	42+5+10 0/10+10/ 10	42+10+0+10 /5	42+10+1 00/5+0	54(U1C E2)
U2CE3 U3CE1 (41)	41+5+10 0/10+0	41+10+100/ 5+10/10	41+10+0 +10/10	52(U1C E3)
U2CE3 U3CE2 (41)	42+5+10 0/10+10/ 10	42+10+100/ 5+0	42+10+0 +10/5	54(U1C E3)

Step 6. Turn to set № 2. Repeat steps 2-4.

U4CE1: 100k / 100k = 1;

U4CE2: 100k / 50k = 2;

U4CE3: 100k / 50k = 2;

In table 4 we present only the results of calculations.

Table 4. The objective function value for each combination of set №3

	a (28)	b (28)	c (52)	d (54)	e (52)	f (54)	Min()
U4CE1 (1)	28+ 1+0 +50 /10	28+ 1+0 +50 /5	52+ 1+2 00/1 0+0	54+1 +200/ 10+5 0/5	52+1 +200/ 10+0	54+1 +200/ 10+5 0/10	34(a)
U4CE2 (2)	28+ 2+2 00/1 0+0	28+ 2+2 00/1 0+5 0/5	52+ 2+0 +50 /10	54+2 +0+5 0/5	52+2 +200/ 5+50/ 10	54+2 +200/ 5+0	59(в)
U4CE3 (2)	28+ 2+2 00/1 0+5 0/5	28+ 2+2 00/1 0+0	52+ 2+2 00/5 +50 /10	54+2 +200/ 5+0	52+2 +0+5 0/10	54+2 +0+5 0/5	59(д)

Step 7. Select the minimum value of objective function for block 3.

According to calculations

min f(x) = 34 for (U4CE1, U2CE1U3CE2, U1CE1).

V. The Strategy Of Input Queue Of Tasks Processing

We have considered the issues of planning a separate task represented as a workflow at an affordable set of heterogeneous resources. In this section, we will discuss approaches to processing and scheduling of various types of tasks coming into a single input queue of workflow management system.

In [13] the following existing multiple workflow scheduling strategies are presented:

1) Scheduling and execution DAGS that are in the input queue one after another. The disadvantage of this approach is the ineffective grid resources utilization, inability to reflect the priorities and the required level of QoS.
2) Scheduling and execution of DAGs in accordance with the criterion of total estimated runtime. Processing order may be different: the priority for tasks with a minimum execution time or the maximum. Such approach does not solve the problem of effective resources utilization and QoS considering.
3) Combining multiple DAGs into a single DAG with a further usage of existing methods of single workflow task scheduling in a heterogeneous environment.

The four main approaches of merging DAGS are determined:

1) Combining DAGS by adding a new entry and new exit "empty" nodes (C1);
2) A composite graph is created in the same way as before, but the scheduling is made by the levels for independent parallel tasks (level-based ordering) (C2);
3) Scheduling and execution of different computational units of workflow tasks occur in the style of round-robin: if on the previous step the task of one workflow was planned and carried out, then on the next step it will be considered as the ready task of another workflow (C3).
4) When combining DAGs into a single workflow structure the estimated execution time of workflow is taken into account and merging by introducing additional nodes occurs at the appropriate level (C4).

Two fairness policies of resources distribution based on calculating the delay of each workflow while choosing the next workflow for scheduling have been introduced in [13].

However the merging DAGs approaches and fairness policies can be applied only to the workflows with the same priorities.

In [14] a ServeOnTime strategy is proposed and its efficiency in comparison with the classical approach FCFS is shown. The strategy is based on adding the new arrived workflow task to the exiting task of executing workflow. Such an approach ignores the QoS requirements, and underutilized resources associated with the occurrence of "gaps" (waiting for completion of the tasks of the previous level and data transfer). In [15] a GapSearchScheduling algorithm is presented. The algorithm is based on finding and filling such gaps by tasks, the execution time of which is less than the gap size.

In [16] the input queue deadline coordinator structure is presented. The deadline driven (DD) coordinator orders DAGs considering deadlines specified by users. DAG with earlier deadline is processed first. DAG priority is computed as inversely to deadline value. The DD-coordinator should verify that the deadline is realistic.

However, the solution is not complete. Deadline set by the user must be considered in relation to the estimated execution time and the arrival time of all tasks.

We suggest the following scheme of tasks priority evaluation:

$$P = U_p + 1/t_{in} \quad (6)$$

where U_p – user task priority set by the policy of appropriate virtual organization;

t_{in} –arrival time of task queued for execution.

When sending a task to perform, the user can set the desired values for the following QoS parameters: restriction to the task time execution (deadline), cost limit of computing (maximal cost), as well as the significance of the coefficients of these parameters.

Taking into account the dynamicity of the grid environment structure, compliance with user-defined QoS-parameters can not be guaranteed, but finding the optimal solutions based on established significance coefficients is guaranteed. Defining the actual values of QoS parameters is possible by the use of simulation model of task execution process.

Guaranteed compliance deadline is possible only for the tasks submitted with advanced reservation policy, and the preliminary assessment time regarding to the task execution time is set by the workflow management system administrator. In the case of low QoS level of available resources replication approach can be used in addition.

Internal xml task specification format has been developed. The format allows describing the tasks of different types, as well as to determine the volume of data transferred between the computational units of workflow task and periodic synchronization between the parallel blocks. When the task is sent to a specific computing resource the task format is converted to those required by the corresponding middleware.

There are static and dynamic approaches to task scheduling. Static approach assumes the availability of information about the current state of grid-network resources and sequence blocks execute tasks prior to computation. Dynamic approach takes into account the dynamic grid-network resources, as well as handles branching in the structure of the problem. However, this greatly complicates the planning process.

The paper introduces the use of a hybrid approach to planning, which is to use static methods for primary distribution, followed by the dynamic regulation of the primary distribution, taking into account the dynamics of the task and the state of network resources. Such a scheme is implemented at the level of the metascheduler through periodic survey of the state of the network resources, control units perform tasks and rescheduling tasks when needed.

The system is supposed to have the following task queues.

1) Single Block And Data Parallel Queue – contains the tasks of the first and the second type. In case of the resource failure, the task is resubmitted to the same queue with the highest priority.
2) Workflow Queue – contains workflow tasks.
3) Workflow Tasks Rescheduling Queue – contains computational units of different workflow tasks requiring rescheduling. Computing unit of any workflow is put into this queue if the resource that was scheduled for the unit failed.

Queue processing and tasks scheduling is made in the order of their priorities. Rescheduling is processed first, and the rescheduling task is assigned to a suitable free resource or the nearest "gap" in the schedule of resource employment.

If several workflow tasks have the same priority, they are merged into a single DAG according to policy C4.

Single DAG scheduling is made using the method presented in Section 5, with further drafting task's schedule taking into account the synchronization between units and graphics of resources employment.

While scheduling the workflow if the amount of the free resources is less than the width of the DAG then the estimated execution time on the set of available resources is compared with the estimated execution time on the greatest possible variety of resources. If the difference in execution time is less than the waiting time of deallocation, the task is assigned to the available resources, or the task is waiting for the release of resources employed and the optimal schedule for its implementation since the liberation of resources is prepared.

The tasks of the first and second type are assigned either to a suitable free resource, or to the nearest appropriate "gap" in the graph of the resource. "Gap" is considered appropriate if the estimated execution time of the task is less than gap size of not more than 80%.

VI. Scheduling Algorithms Effectiveness Evaluation

The experiments were carried out only for the analysis of the effectiveness of the proposed single DAG scheduling method on the available set of resources. Effectiveness evaluation of the proposed strategy for processing the queue of tasks taking into account the dynamics of their implementation requires additional research.

To investigate the properties of scheduling algorithms the GridSim toolkit [17] was expanded by adding new entities required for modeling the processes of planning and execution of workflows in the grid-environment. Implemented modules class diagrams are shown in Fig. 4-5.

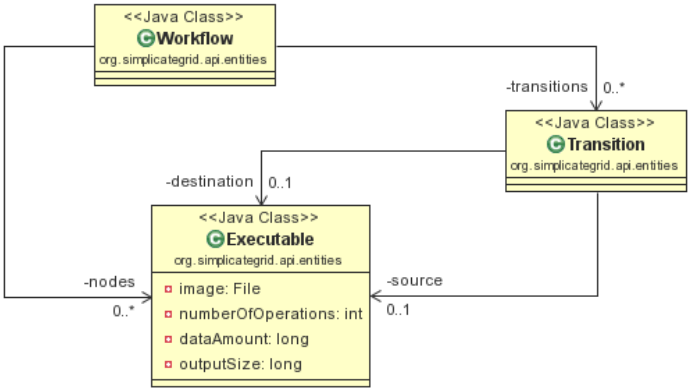


Fig. 4 the class diagram of the workflow specification module

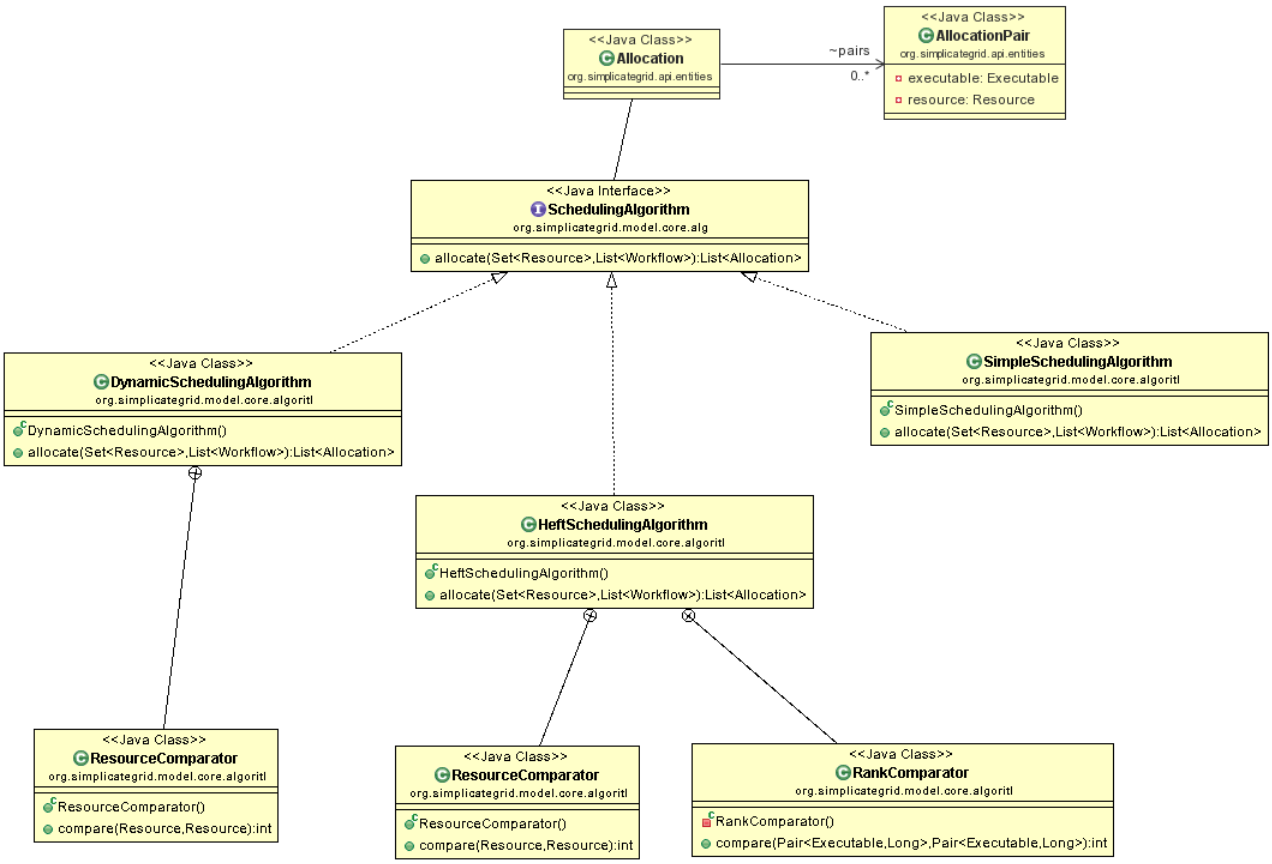


Fig. 5 the class diagram of the metascheduler module

The experiments were carried out for randomly generated workflow tasks of varying complexity. The example of randomly generated graph is shown in Fig. 6.

Obviously, the real workflows of specific domains have fewer nodes and links. However, the use of more complex workflows allows identifying bottlenecks in the studied algorithms.

The grid-environment experimental model was presented by four compute nodes with the following characteristics: CE1 = {10 mips, 100, 0, {100, 50}}; CE2 = {25 mips, 100, 0, {100, 20}}; CE3 = {35 mips, 100, 0, {80, 40}}; CE4 = {47 mips, 100, 0, {100, 30}}.

The effectiveness of a scheduling algorithm in the experiments was determined by: 1) the objective function value; 2) the computational complexity of the algorithm.

In order to simplify the objective function was defined by task execution time, excluding the economic costs.

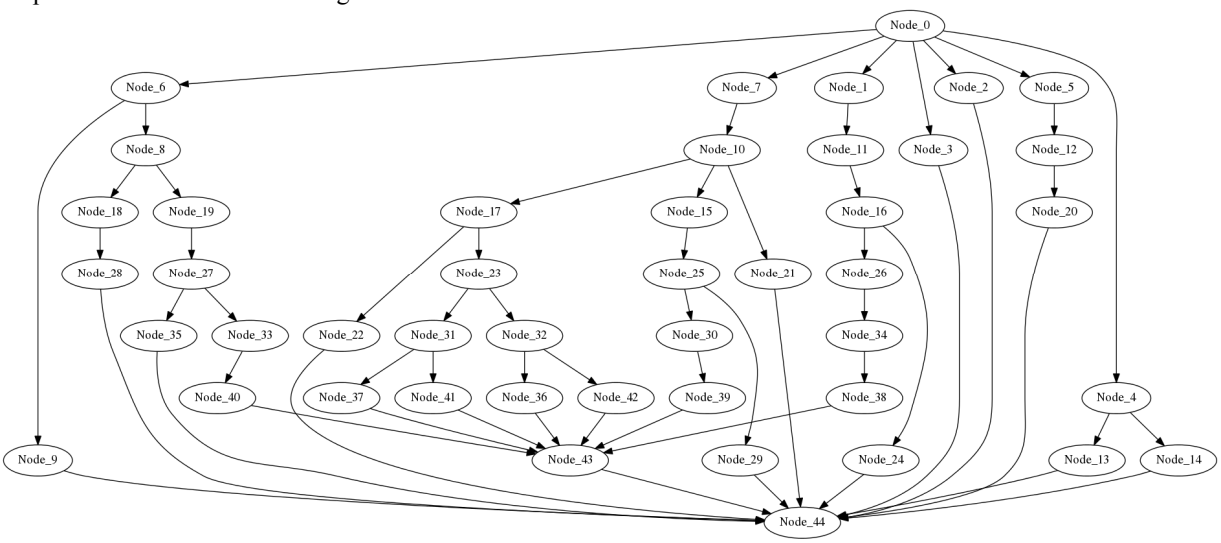


Fig. 6 generated workflow structure example

Table 5 shows the results of the experiments for the tasks of different complexity and the following scheduling algorithms:

1 - heuristic algorithm HEFT, 2 - scheduling algorithm based on dynamic programming method, 3 - random selection of the resource to perform the task of computing unit, ready to run, excluding the cost of accommodation.

Table 5. Experimental results

Number of nodes	Number of links	Number of levels	Scheduling time, ms	Runtime, s	Algorithm	The ratio of the (number of nodes / number of links)
10	13	5	3	518,4	1	0,769231
25	34	7	9	6859,91	1	0,735294
50	72	9	30	7241,28	1	0,694444
10	13	5	5	408	2	0,769231
25	34	7	12	5833,44	2	0,735294
50	72	9	43	6167,76	2	0,694444
10	13	5	1	576	3	0,769231
25	34	7	5	7873,92	3	0,735294
50	72	9	20	8203,2	3	0,694444

Fig. 7-8 shows the results of the algorithms performance criteria evaluation depending on the complexity of the workflow structure. Proposed algorithm showed a higher efficiency for the workflow runtime criterion. Scheduling time of the proposed method is higher than for other algorithms, but it is incomparably less the workflow runtime in the grid-environment that justifies the appropriateness of the proposed solution.

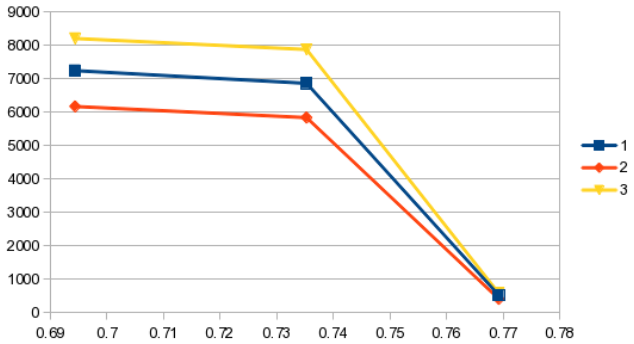


Fig. 7 execution time depending to the complexity of the workflow structure

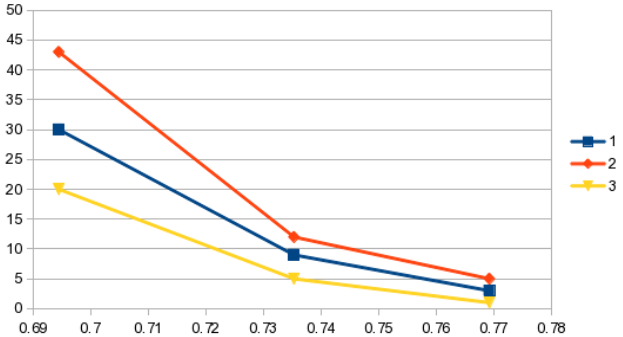


Fig. 8 scheduling time depending to the complexity of the workflow structure

VII. Conclusion

The paper discusses the features of using grid environment to perform various types of computational tasks of high dimension. The metascheduler structure of the centralized workflow management system and two stage scheduling strategy that takes into account QoS requirements, the dynamicity execution and structural features of the task were proposed. The dynamic programming method application to the workflow scheduling problem was shown in the paper. Experimental results proved the effectiveness of the proposed workflow scheduling method. The effectiveness evaluation of the proposed queue processing and rescheduling strategy where not presented in the paper and require additional research.

References

[1] Forti A. DAG Scheduling for grid computing systems / A. Forti // Ph.D. Thesis, University of Udine, Department of Mathematics and Computer Science. – Italy, 2005 – 2006. – P. 43-46, 52-55.

[2] Kazymyr V. Grid workflow design and management system / V. Kazymyr, O. Prila, V. Rudyi // International Journal “Information Technologies & Knowledge”. – 2013. – Vol. 7, N 3. – P. 241 – 255.

[3] Gerasoulis A. A comparison of clustering heuristics for scheduling directed acyclic graphs on multiprocessors / A. Gerasoulis, T. Yang // Journal of Parallel and Distributed Computing. – 1992. – N 16. – P. 276 – 291.

[4] Tompkins M.F. Optimization techniques for task allocation and scheduling in distributed multi-agent operations / M.F Tompkins // Diss. Massachusetts Institute of Technology. – 2003. – P. 20-23.

[5] Job Submission Description Language (JSDL) Specification, Version 1.0, GFD-R.136. – 2008. – P. 5-10.

[6] Extended Resource Specification Language, Reference Manual for ARC versions 0.8 and above, Nordugrid-Manual-4. – 2013. – P. 13-28.

[7] Job description language attributes specification for the gLite Workload Management System, WMS-JDL.doc. – 2011. – P.7-10, 38-40.

[8] Yu J. Workflow Scheduling Algorithms for Grid Computing, Metaheuristics for Scheduling in Distributed Computing Environments / Yu J., Buyya R., Ramamohanarao K.; Xhafa F., Abraham A. (Ed.). – Berlin, Germany: Springer, 2008. – P. 111-149.

[9] Fangpeng D. Scheduling Algorithms for Grid Computing: State of the Art and Open Problems / D. Fangpeng, Akl.G. Selim // School of Computing, Queen’s University Kingston, Ontario, Technical Report. – 2006. – N 504. – P. 7-32.

[10] Liou J. CASS: an efficient task management system for distributed memory architectures / J. Liou, M.A. Palis // International Symposium on Parallel Architectures, Algorithms and Networks (ISPAN '97). - Taipei, Taiwan, 1997. – P 289 – 295.

[11] Bajaj R. Improving Scheduling of Tasks in a Heterogeneous Environment / R. Bajaj, D.P. Agrawal // IEEE Transactions on Parallel and Distributed Systems. – 2004. – Vol. 15, N 2. – P. 107 – 118.
[12] Introduction to algorithms, third ed. / Thomas H.C., Leiserson C.E., Ronald L.R. [et al.]. – [3 ed.]. – Cambridge, Massachusetts London, England: The MIT Press, 2009. – P. 357 – 414.
[13] H. Zhao, R. Sakellariou: Scheduling Multiple DAGs onto Heterogeneous Systems. Proceedings of the 20th international conference on Parallel and distributed processing, p.159-159, April 25-29, 2006.
[14] L. Zhu, Z. Sun, W. Guo, Y. Jin, W. Sun, W. Hu: Dynamic Multi DAG Scheduling Algorithm for Optical Grid Environment. Network Architectures, Management, and Applications, V 6784(1), 2007.
[15] Bittencourt, L.F., Madeira, E.R.M.: Towards the Scheduling of Multiple Workflows on Computational Grids. Journal of Grid Computing, 8, pp. 419–441, 2010.
[16] Melnyk A. Multiple DAGs Scheduling with Deadline Driven Coordinator in Grid / A. Melnyk // Second International Conference "Cluster Computing". – Lviv, Ukraine, 2013. – June 3–5. – P. 127 – 130.
[17] Buyya R. GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing / R. Buyya, M. Manzur // The Journal of Concurrency and Computation: Practice and Experience (CCPE). – 2002. – Vol. 14, Is. 13–15. – P. 1179 – 1219.

ROI Sensitive Analysis for Real Time Gender Classification

Marcos A Rodrigues, Mariza Kormann and Peter Tomek

***Abstract*—**This paper addresses the issue of real time gender classification through texture analysis. The purpose is to perform sensitivity analysis over a number of ROI-Regions of Interest defined over face images. The determination of the smaller ROI yielding robust classification results will be used for fast computation of texture parameters allowing gender classification to operate in real-time. Results demonstrate that the ROI comprising the front and the region of the eyes is the most reliable achieving classification accuracy of 88% for both male and female subjects using raw data and non-optimised extraction and classification algorithms. This is a significant result that will drive future research on optimisation of texture extraction and linear discriminant algorithms.

***Keywords*—**Face detection and tracking, texture analysis, gender classification

I. Introduction

IMAGE analysis for gender classification has a number of useful applications such as collecting demographic statistics for marketing purposes, security surveillance, and the development of customised human-computer interfaces. The ADMOS project [1] is funded by the EC and aims to develop a real-time gender and age recognition to be used in private spaces of public use, such as shopping malls, fairs and outdoor events.

The purpose of this paper is to report on sensitivity analysis and performance evaluation concerning gender classification using texture analysis. The parameters under investigation include the size of the detected region in the image in relation to the location of the eyes, and various regions of the face. Specifically these include a larger image with aspect ratio *width*:*height* of $3:4$ which normally includes ears, hair and a portion of the neck, $1:1$ which is narrowed down to the region of the face from the front to the chin and normally includes portions of the ears, and $2:1$ which are defined as the top half of the face (front, eyes, portions of the nose and ears) and bottom half (including lips and portions of nose, chin and ears).

A set of experiments are designed to determine the optimal region of the face; ideally this would be as small as possible to allow the system to operate in real time. The techniques under investigation are the LBP–Linear Binary Pattern algorithm [2] in conjunction with eigenvector decomposition to determine class membership [3]. No LBP improvements are proposed here neither the optimisation of classification algorithms; the aim is to robustly assess the various ROI-Regions of Interest of an image. Once such regions are determined, then improvements to the techniques are investigated.

This project has received funding from the European Union Seventh Framework Programme for research, technological development and demonstration under grant agreement number 315525, 2013–2015.

Marcos A Rodrigues and Mariza Korman are with the GMPR–Geometric Modelling and Pattern Recognition Research Group at Sheffield Hallam University, Sheffield, UK. Email {m.rodrigues, m.kormann}@shu.ac.uk

Peter Tomek is with ATEKNEA, Budapest, Hungary. Email peter.tomek@ateknea.com

LBP is a non-parametric method used to summarise local structures of an image and have been extensively been exploited in face analysis for gender, age, and face recognition [4], [5], [6], [7], [8], [9]. Normally, LBP are employed in local and holistic approaches and a number of extensions have been demonstrated in the literature (e.g. [6]) in connection with linear discriminant analysis and SVM-support vector machines.

The approach in this paper is demonstrated by using a public database from which the various regions are automatically selected by face, eyes and lip detection algorithms. The method is described in Section II, experimental results are presented in Section III with a conclusion in Section IV.

II. Method

In order to perform real-time gender classification, a number of steps are necessary. First faces must be detected in the image and this is achieved through the Viola-Jones method [10], [11] available from OpenCV libraries. An unconstrained image may contain a number of faces and each region of interest ROI containing a face must be processed independently and the detected gender must be assigned to a corresponding data structure (to that region of interest). The data structure thus, will contain a gender attribute such that to avoid unnecessary multiple calls to the gender classification function if a particular tracked face has already a gender definition. This applies to the tracking of faces over multiple frames, but these aspects are not discussed further in this paper. Instead the focus is on sensitivity analysis over selected regions of the face.

Sensitivity analysis starts with testing aspect ratios of the ROI concerning the selected facial region. The Viola-Jones method yields a ROI with aspect ratio of $1:1$. The method, however, suffers from false positive detection: due to the simplicity of the Haar-like features used, some of the detected face objects are not faces at all. In order to verify whether or not it is a face, a number of constraints are imposed namely, a face must have two eyes and a lip. To verify the constraints each ROI is taken in turn. First eye detection is performed in the knowledge that right and left eyes are located in the first and second quadrants of the ROI, and lip detection in the knowledge that the lip is spread over the third and fourth quadrants. An example of a verified face is depicted in Figure 1. If the ROI satisfies these constraints then proceed to gender classification. Note that eye detection is a specific requirement of the ADMOS project but very useful in the context of sensitivity analysis presented here as the size of a larger ROI (of aspect ratio of 3:4) is determined from the eye locations.

Local binary patterns [2], [4] are grey-scale operators useful for texture classification defined over local neighbourhood pixels. It was originally defined using a 3×3 array of pixels. The value of the centre pixel is compared with its neighbours

ISBN: 978-1-61804-236-1

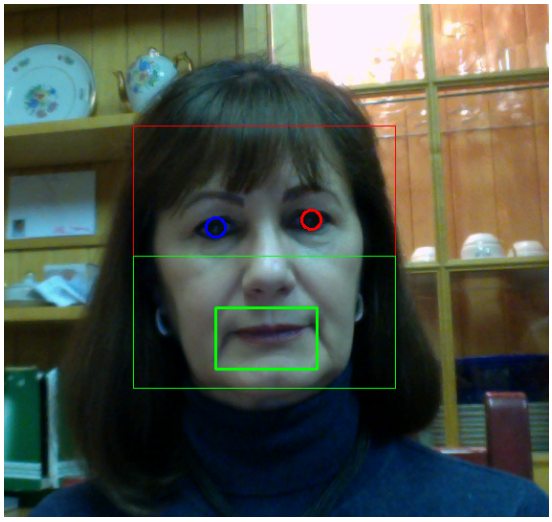


Fig. 1. Face, eyes, and lip detection in real time with annotated face ROI of aspect ratio 1 : 1

and the result (greater or smaller) expressed as a binary number and summed over all pixels considered. LBP can be expressed over P sampling points on a circle of radius R where the value of the centre pixel (x, y) is expressed as:

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} T(I_p - I_c)2^p, \tag{1}$$

where I_p and I_c refer to the pixel intensity in the grey level of the centre pixel and of P pixels on a circle of radius R, and T is a thresholding function with $T(.) = 1$ if $(I_p - I_c) \geq 0$ or $T(.) = 0$ otherwise. Normally images are defined in blocks from which individual LBPs are calculated and then concatenated into a single histogram. The analysis of such histograms can be used to differentiate texture patterns. The size of the block under analysis can vary and this obviously will be reflected in the LBP histogram.

In order to improve the robustness and discriminative power of the basic LBP operator as defined in equation (1) a number of variations to LBP have been proposed in the literature [5]. Note that the purpose of this paper is not to investigate possible extensions to LBP but to perform a sensitivity analysis to determine which region of the face is the most reliable for gender classification. Once this is achieved the most promising region will be investigated with a number of extensions to LBP and linear discriminant algorithms [3] including perceptron, relaxation rules, Fisher's criterion, least mean squared methods, and support vector machines as reported in the literature (e.g. [5], [6], [7]).

The approach to gender classification applied to facial regions can thus be stated as:

1) Define a set of stable measures $m_i (i = 1, 2, \ldots, n)$ on an image ROI and build a vector $M = (m_1, m_2, \ldots, m_n)^T$ that characterises the ROI;
2) Build a matrix Ω of vectors M where the index of M points to the identity of the ROI object: $\Omega = (M_1, M_2, \ldots, M_s)^T$ where s is the total number of images or vectors in the database;
3) Define a method to estimate the closest distance from a given ROI vector M to the most similar vectors in the database. The class of those vectors will point to the most likely class of M.

The set of stable measures is the histogram produced as a result of applying LBP to the data. It subsumes a large number of variables on the image ROI and the purpose is to identify which variables are more relevant to the problem at hand. Here the problem is approached by analysis of variance, or principal values. There are at least three ways a set of principal components can be derived. If m_i is the set of original variables, then they can be expressed as a linear combination ψ:

$$\psi_i = \sum_{j=1}^{p} a_{ij} m_j \tag{2}$$

The Hotelling approach described in [3] is used here in which the purpose is to find a linear separation for which the sum of the squares of perpendicular distances is a maximum, that is the choice is to maximise the variance of ψ. This is a common approach to the problem based on first determining the scatter matrix S by subtracting all values from their mean. Then the covariance matrix is estimated as $\Sigma = SS^T$. The principal components are determined by performing an eigenvector decomposition of the symmetric positive definite matrix Σ and then using the eigenvectors as coefficients in a linear combination of the original variables m.

In order to determine class membership, the method is to define a training set in which ground truth representative of male and female images are used and classified accordingly. The leave-one-out technique is used at testing stage. The principle is that the closest vector in the database real class is the proposed class for the testing vector. In order to improve accuracy, a voting method is proposed as the best of 5 closest matches – that is, for an unknown vector the proposed class is determined by majority counting.

III. Experimental Results

A public database is used containing a large variation of faces concerning pose and illumination, details can be found in [12]. The database provides sets of large images which are not restricted to the facial area, and subjects are presented in a number of different poses other than frontal. Examples of images from the database are depicted in Figure 2.

A set of five experimental tests were carried out as follows.

1) ROI1: using a large ROI that includes hair, ears, and areas of the neck. This is selected from face detection and by adjusting the width and height of ROI with aspect ratio 3:4 based on the detected distance between eyes. For instance, if the inter-ocular distance is d, then the ROI width is defined with d to either side of the eyes, and $2d$ to the top and $2d$ to the bottom.
2) ROI2: using a ROI of 1:1 as detected by the face detection algorithms. This is narrowed down to the area of the face as compared to the previous larger ROI, and normally includes the areas from the front to the chin.
3) ROI3: using a ROI of 2:1 including only the top half of the detected face.
4) ROI4: using a ROI of 2:1 including only the bottom half of the detected face.
5) ROI5: using a ROI of 2:1 corresponding to the detected area around the lips.

A set of 100 images were used, 50 male and 50 female. For each ROI, LBP was performed to extract the relevant features represented by the histogram. An example is shown in Figure 3. The histograms for each ROI are used for training

ISBN: 978-1-61804-236-1

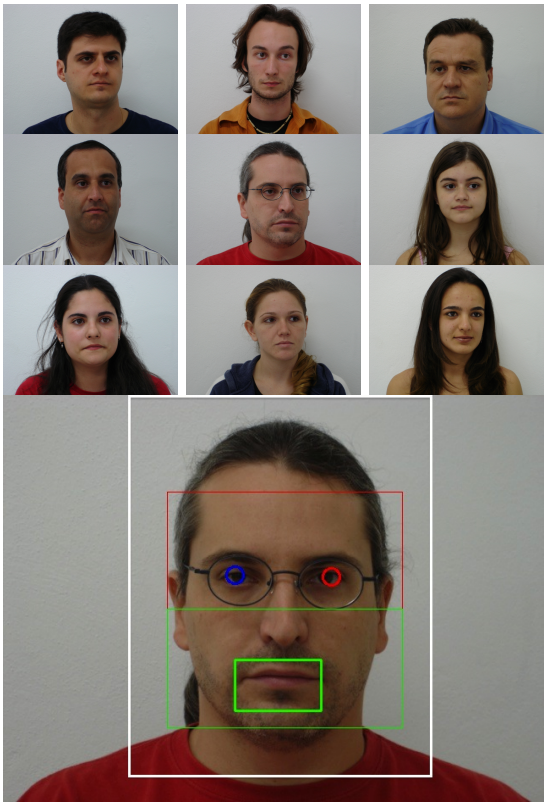


Fig. 2. Examples of data from the FEI database. The larger image shows the 5 automatically detected regions of interest.

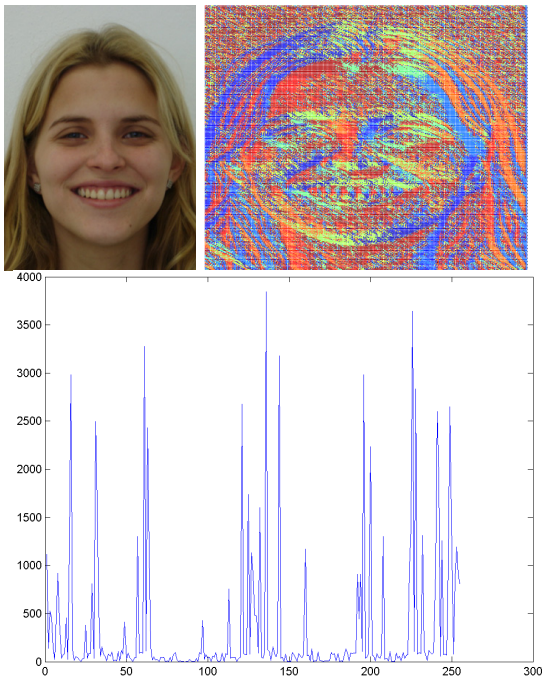


Fig. 3. Example of LBP and histogram for a selected ROI.

and testing purposes. Each ROI was tested independently as the purpose is to determine which one yields more accurate results; specifically the smaller effective ROI is sought as the purpose is to be applied in real time gender classification.

The implementation was carried out in Matlab following the method described in Section II. Results were automatically saved to a spreadsheet for further analysis. Performance in terms of accuracy were noted and results are summarised in Table I. Results for ROI1 yielded the highest accuracy for male classification at 92%, and female classification was substantially lower at 79%. This is probably because adding portions of background together with large amounts of hair increases the overall variance which can be confused with skin variations of male subjects due to facial hair. In any case, the inclusion of background is not desirable as it can introduce errors that cannot be controlled.

Results for ROI2 were slightly improved for female subjects at 83% but worst for male ones at 83%. Still, results are more consistent and thus, preferable to ROI1. ROI3 (consisting of the upper half of ROI2) yielded the best results, at 88% accuracy for both male and female subjects. This shows that there is enough variation in this region to detect both classes with good accuracy. ROI4 (consisting of the bottom half of ROI2) retained its previous accuracy for male subjects at 88% but for female the performance decreased substantially to 71%. This shows that the region does not contain enough variation in texture to guarantee robust classification, which is somewhat surprising. Before testing this ROI, it was expected that, due to facial hair and skin texture variations on either shaved and unshaved faces, male and female subjects would be detected with the highest degree of accuracy. Finally ROI5 has proved to be problematic as automatic lip detection failed in about half of the data (multiple lip detection was deemed a failure). Thus the lack of robust lip detection rendered this ROI unstable and the statistics reported here at less than 40% are just an indication of performance.

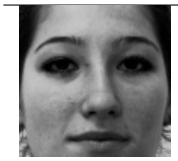
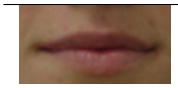
Since this research is only interested in the minimum ROI with highest performance, results demonstrate that ROI3 is the most promising one. The high variance caused by skin texture, eyebrows, eyes, and portions of hair mean that the region can be further exploited for more robust classification. With this information, it is now possible to focus on improvements to LBP and to the classification algorithms. The obvious avenues to explore include the use the Fisher's criterion and SVM-support vector machines. In particular, SVM holds the highest promise as it is a technique designed to maximising the margin between canonical hyperplanes separating the two classes. With a higher margin, the probability of misclassification naturally decreases and it should be possible to achieve higher accuracy for both male and female subjects.

IV. Conclusion

This paper has presented a sensitivity analysis of gender classification based on LBP and eigenvector decomposition over 5 regions of interest. The adopted methodology was directed towards testing the LBP algorithm over the selected regions to determine the smaller region of interest yielding the best classification results. Images from a public database were used on which automatic detection of the face region, eyes, and lips were performed resulting in the various regions of interest being defined. Classification results show that the region containing parts of the nose, eyes and front are the most

ISBN: 978-1-61804-236-1

TABLE I
COMPARATIVE ANALYSIS OF IMAGE REGIONS

Image ROI	Gender	Classification results
 ROI1	Male Female	92% 79%
 ROI2	Male Female	83% 83%
 ROI3	Male Female	**88%** **88%**
 ROI4	Male Female	88% 71%
 ROI5	Male Female	<40% <40%

reliable, with an overall accuracy of 88% for both male and female subjects.

This is a significant result on its own right but, more importantly, it provides a focus on which to apply optimisation techniques to bring overall accuracy to a desirable level of 95%. Other techniques reported in the literature such as in [6] perform specific selection of histogram bins and use SVM for feature classification and are shown to perform slightly better than using the raw LBP histogram as reported here. Those results and the results obtained in this paper clearly indicate that, by using the selected region and developing similar techniques it will be possible to substantially increase classification accuracy. Furthermore, by allowing the use of a substantially smaller region as compared to the ones reported in the literature, computation will be much faster and appropriate for real-time applications. Research is under way on fine-tuning LBP for the selected region through the use of SVM and related methods and results will be reported in the near future.

REFERENCES

[1] ADMOS (2013) *Advertising Monitoring System Development for Outdoor Media Analytics*, EC Grant Agreement 31552. [Online] Available at http://admos.eu

[2] T. Ahonen, A. Hadid, and M. Pietikäinen. Face description with local binary patterns: Application to face recognition. *TPAMI*, 28(12):20372041, 2006.

[3] A. Webb and K. Copsey (2011) *Statistical Pattern Recognition*, 3rd edition, Wiley, 666pp.

[4] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen. (2011) Computer Vision Using Local Binary Patterns. *Springer*.

[5] Ylioinas, J., Hadid, A., Pietikäinen, M. (2011) Combining contrast and local binary patterns for gender classification. *SCIA 17th Scandinavian Conference on Image Analysis.*

[6] C. Shan (2012) Learning local binary patterns for gender classification on real-world face images, Patter Recognition Letters 33 (2012) 431–437.

[7] Y. Guo, G. Zhao, M. Pietikäinen, and Z. Xu. Descriptor learning based on fisher separation criterion for texture classification. *In Proc. ACCV10*, 185–198, 2010.

[8] Lian, H., Lu, B. (2007) Multi-view gender classification using multi-resolution local binary patterns and support vector machines. *Int J Neural Systems* 17 (6), 479–487

[9] Sun, N., Zheng, W., Sun, C., Zou, C., Zhao, L. (2006) Gender classification based on boosting local binary pattern. *In: Int Symp on Neural Networks.*

[10] Viola, P., Jones, M. (2001) Rapid object detection using a boosted cascade of simple features. *In: IEEE Conf. on Computer Vision and Pattern Recognition* (CVPR), pp. 511518.

[11] Viola, P., Jones, M. (2004) Robust real-time face detection. *Internat. J. Comput. Vision* 57 (2), 137154.

[12] (2014) FEI Face Database. [Online] Available http://fei.edu.br/ cet/facedatabase.html

Analysis of new collaborative writing within Web 2.0

P. Cutugno, L. Marconi, G.Morgavi, D. Chiarella, M. Morando

***Abstract*—**In recent years, the transition from Web 1.0 to Web 2.0 enabled the creation of content by the users of the Network: social networks, blogs, forums, chats and wikis have arisen..

Phenomena, such as collaborative/collective writing, already born at the beginning of the 20th century, found their natural setting, a wide audience of reference of "writers' and readers in multiple languages within the Web 2.0.

In this paper our goal is to verify if and how the characteristics of the textual analysis of narrative plots can be used for the analysis of collaborative narrative texts. In particular, we will check if features like correctness, completeness, consistency and coherence together with tools for statistical analysis of language suitable for analysing the new collaborative writing 2.0.

***Keywords*—**Collaborative Writing, Social Publishing, Textual Analysis, Social Implications of Modern Communications.

I. INTRODUCTION

The term collaborative writing refers to projects where multiple people together (collaboratively) create written works rather than individually. In a true collaborative environment, each contributor has an almost equal ability to add, edit, and remove text. The writing process becomes a recursive task, where each change prompts others to make more changes.When a group of authors shares creative control of a story the form of writing is called collaborative fiction.Collaborative writing is easier to do if the group has a specific final goal in mind and harder if the goal is absent or vague.Writing games for collaborative writing have a tradition in literary groups such as the Dadaists.

There are several of possible degrees of collaboration in authoring. At one end of the range there is a single author who, through discussion with and review by colleagues, produces a document. In the other end of the spectrum there is a group of writers who jointly author a document. Lowry et al.[1]identified five coordination strategies for group writing: single-author writing, sequential single writing, parallel writing, reactive writing and mixed mode. Each strategy has inherent advantages and disadvantages.Traditional fiction writers and writing circles have experimented in creating group stories. Collaborative writing can greatly increase motivation and speed of production for authors[2].

Extending the three types of co-authoring described by Ritchie and Rigano[3] collaborativewriting can be organised as:

- Turn writing. In this form of writing, authors contribute different sections of a text which are then merged and harmonized by a lead author.
- Lead writing. One person drafts the text, which is amended by the others.Alternatively, authors might write the text for their own particular subplot within an overall narrative, in which case one author may have the responsibility of integrating the story as a whole.
- Writing together side-by-side. A text is composed by two or more persons who think aloud together, negotiating and refining the content. A collaborative author may focus on a specific protagonist or character in the narrative thread, and then pass the story to another writer for further additions or a change in focus to a different protagonist.

Italy has a strong traditionin collaborative fiction: one of the oldest and the most remarkable texts was*Lo zar non è morto*, a 1929 collective novel by the futurist team "Gruppo dei Dieci".

In the 'Letture Americane' Italo Calvino[4]defines the hyper -fiction as a place ' of contemporary infinite universes where all possibilities are realized in all possible combinations'. It is amazing, at a time when today's technological developments were unpredictable, that Calvino could imagine to build a hyper -fiction content. In 1941 Luis Borges[5] wrote, "The Garden of Forking Paths" a novel in which "describes every possible outcomes of an event, each of which leads to a further multiplication of consequences in a continuous branching of possible futures". This is an example of hypertext novel realized on paper. The first hypertext novel written for the Network was '*Afternoon, a story'*by American author Michael Joyce in 1987[6]. It was published by Eastgate Systems in 1990. Its hypertext structure, its links to other nodes of the novel, the chance to choose the plot more pleasing to the reader, earned him some attention from the web surfers and generated both commercial and critical success.

The advent of the Internet has seen many such collaborative writing experiences online, resulting both in hypertext fiction and in more conventional literary production.

In more recent years(1994) in Bolognaa number of cultural activists (hackers, writers and cultural workers)began using the pseudonymLuther Blissett (the name of the first black footballers to play in Italy[7]), for staging a series of urban and media pranks and to experiment with new forms of authorship and identity. From Italy this open-identity multiple-use name spread to other European cities, such as London, as

ISBN: 978-1-61804-236-1

well as countries such as Germany, Spain, Sloveniaand also Canada, United States, and Brazil. According to Marco Deseriis[8], the main purpose of the Luther Blissett Project (LBP) was to create "a folk hero of the information society" whereby knowledge workers and immaterial workers could organize and recognize themselves. Thus, rather than being understood only as a media prankster and culture jammer, Luther Blissett became a positive mythic figure that was supposed to embody the very process of community and cross-media storytelling. Roberto Bui[9], one of the co-founders of the LBP,explains the function of Luther Blissett and other radical folk heroes as mythmaking or mythopoiesis:"Mythopoiesis is the social process of constructing myths, by which we do notmean 'false stories', we mean stories that are told and shared, re-told and manipulated, by a vast and multifarious community, stories that may give shape to some kind of ritual, some sense of continuity between what we do and what other people did in the past. A tradition: in Latin the verb 'tradere' simply meant 'to hand down something', it did not entail any narrow-mindedness, conservatism or forced respect for the past. Revolutions and radical movements have always found and told their own myths".This collective has become famous with the novel Q, a historical novel published in Italy by Einaudi publishing house in multiple languages.

Wu Mingcollective (extended name: Wu Ming Foundation) is thenom de plume for a group of Italian authors stemmed in 2000 from a subset of the "Luther Blissett" community in Bologna. In Chinese, '*wu ming*' can mean "anonymous" or, with a different tone on the first syllable, 'five people'. The name is meant both as a tribute to dissidents ('Wu Ming' is a common byline among Chinese citizens demanding democracy and freedom of speech) and as a rejection of the celebrity-making machine which turns the author into a star.'*wu ming*' is also a reference to the third sentence in the *Dao de jing*: "Nameless is Heaven's and Earth's origin"[10].

Unlike the open name "Luther Blissett", "Wu Ming" stands for a defined group of four known writers active in literature and popular culture. The band authored several novels, some of which have been translated in many countries. Their books are seen as part of a body of literary works (the "nebula", as it is frequently called in Italy) described as the New Italian Epic, a phrase that was proposed by Wu Ming themselves.

Another example of collaborativewriting is represented by fourteen-writers,an internationalcollective from 14 different nations around the world who decided to join their efforts and their narrative skills to create a novel - the first "international" novel (Global Novel) in history of literature-: 'My name is nobody'. All the details of the book, from the story to the writing style, were planned during a single meeting in Athens.

Currently the collective industrial writing and the relay writing are consolidated in the Italian Cyber Network.

The collective industrial writing is the brainchild of two Italians in 2007, driven by the desire to "give birth to a network of readers and writers careful to the innovation and sensitive to the issue of sharing of knowledge. Their goal was that the collective writing of small groups becomesa literary practice. They aim to write, in such a way an open great novel, a collective book written by hundreds of users that should be, first of all, a good book"[11].

The industrial collective writing is based on an approachplanned in every detailby the " artistic director " who formally does not participate in the writing process, but who leads the writing,selects the topics and guidelines of the plot. The task of the artistic director is, at first, to prepare outlinesof writing textsfocused on the characters, the templates and the plot. The members of the writing group will give life to each outline producing their own novels. In the next step the artistic director together withthe collective of writers, will make a collage and editing of novelsproducing the final text of the novel. The idea,underlying thismethodology, is "to overcome the individual approach to the literature in the age of participatory culture more and more spread through the computer network"[12].

Relay writing [13] began in 2008 with the stated aim to give users the freedom to propose content. The current writer is called the guardian of the pen.Each guardian has the chance to decide to leave the pen to the next writer for the subsequent contribution. Only writerswho proposes an outline tuned with the novel plot image foreshadowed by the guardian will be chosen. In turn, the second writer becomesthe next guardian of the pen. Two main ideas underlying this approach: the anarchy of the contents is counteracted and the stiffness of plot and deadlines are avoided for the writing group.

II. COLLABORATIVE WRITING SOCIAL NETWORKS

In 2013 the first collective writing social network introduced a new web 2.0 collaborative writing shape: 20lines[14].

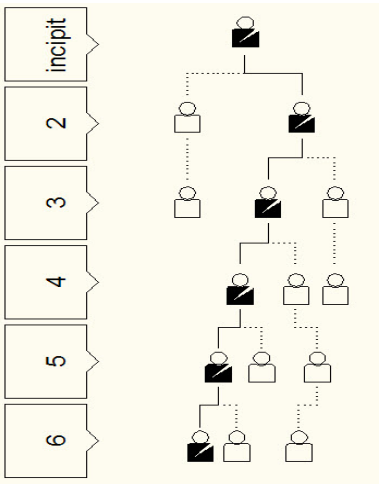


Figure 1 20lines plotting architecture

In this social network, each novel consists of 6 sections of up to 20 lines each, each section can be written by anyone, after registration to the site, so that starting from one incipit endless plots can developed. Readers will assess the best plot through typical social network's interactive instruments such

as rates, shares, etc. . After 20 days of publication on the web site, the plot that received the best rates from readers will be selected for permanent publication on 20lines site. Another typical characteristic of social network is that each writer can associate to its section or incipit, images, videos, maps, creating a hypertext

This social publishing site is areal platform offering opportunitiesforcollective writingof novelswithin a social frame by using instruments such as 'likes', 'views','followers', etc. . This Italian start-up founded in 2013 by three 24-26 years oldmen, with the aim to make the editorial processdemocratic and to use literature as a communication tool.

Herewe do not propose an analysis of the development of narrative plots as proposed by Propp in the "Morphology of the Folktale"[15]. In his analysisPropp identifies specific narrative patterns, in particular, he recognizes that the development of the plot follows a basic skeleton identified by: a balance, a breakdown in the balance, and a restoration of the balance itself. We do not even suggest an analysis which aimsto show the importance of identifying a narrative model like the one done in the 60s by Eco, Barthes, Genette, Todorov, Greimas[16]. In this paper we aim to investigate if and how the characteristics of the textual analysis used in literature novel plots can be applied for the analysis of collaborative narrative texts created in social networks. In particular, we'll check if features like correctness, completeness, consistency and coherence together with a statistical analysis of language tools are suitable for the analysis of the new web 2.0 collaborative writing.

Generally speaking, the analysis of textual datacan be referred to as a qualitative, quantitative or mixed analysis. The process underlying it, however, always refers to the recognition, classification, coding and synthesis of the text[17].

Inthe linguistic analysis a set of procedures are provided, such as:

- the segmentation of the text into sentences;
- the tokenization: the segmentation of the text into words or tokens;
- the morpho-syntactic annotation, where each token found is associated to the information on the grammatical category that the specific word has in its context and to its relative lemma;
- the syntactic annotation that involves the analysis of syntactic structure of the sentence in terms of dependency relationships.

The text analysis of the novels have been processed by "LinguA: Linguistic Annotation pipeline" which combines rule-based and machine learning algorithms, developed by the Institute of Computational Linguistics "Antonio Zampolli"[18][19][20].

The proficiency levels of the Italian language within the novel have been measured through parameters of lexical, morpho-syntactic and syntactic skills. Results are compared with the Basic Vocabulary of the Italian Language (Vocabolario di Base VdB) defined by Tullio De Mauro[21]. TheVdB is a list oflemmas representing the share of the Italian language used and understood by most of Italian speakers.

In particular, the entries of the VdB are classified into three levels:

- Fundamental Vocabulary: the first 1,991 entries are the most frequent in the Italian language;
- Widely used Vocabulary: the subsequent 2,750 entries are still very frequent, although much less than those of the fundamental vocabulary;
- High-availability Vocabulary: its 2,337 lemmas are " relatively uncommon in the written language, but very common in the spoken Italian language.

We analysed two completednovels of 20lines: "Rigor mortis", " Sono stata Brava " (I was good).

III. THE DATA

A. *The novel 'Rigor mortis'*

" Rigor mortis " is a fiction written by sixwomen. Its plot is focused on a trip they took, from south Italy to Cuneo (a Northwestern city), to go,together,to the funeral of aprofessor of the Roman University 'La Sapienza'.Each of them had a short loveaffairwith this teacher when they were students. Our analysis is focused both on the social and the purely linguistic characteristics of the novel.

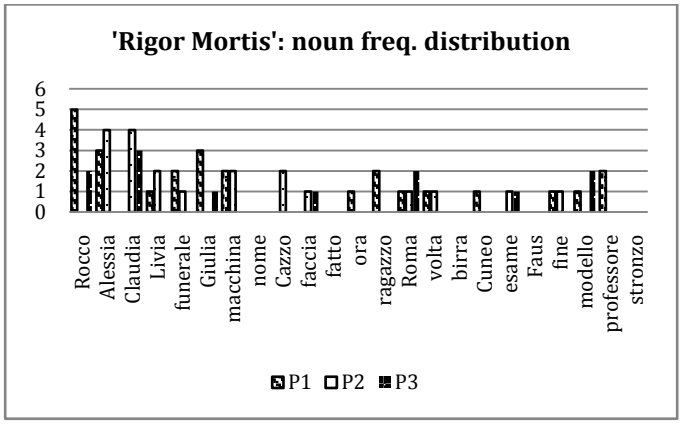


Figure 2 Noun frequency distribution for Levels of plot 1 (P1),2 (P2) and 3 (P3)

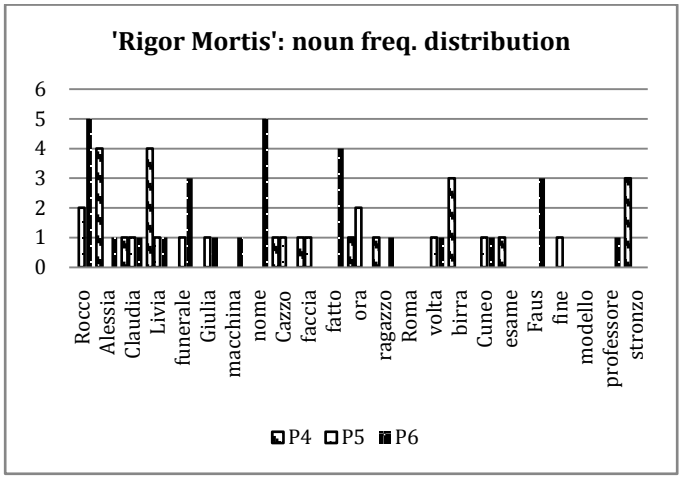


Figure 3Noun frequency distribution for Levels of plot 4 (P4), 5 (P5) and 6 (P6)

A careful analysis of the profiles of the authors and

theircomments to the novel showed thatthe authors have previously planned the plot and the characters. The six writers seem to be very active onthe analysedsocial publishing platform or on Facebook.Some of them are already network users that are very popular on other sites like their personal blogs. Although the novel appears to comply with all the analysedfeatures (correctness, completeness, consistency and coherence) concerning the narrative plots, we must observe that the usage of some images and videos within the 20 lines single author section did not alter in any way the novel plot.The novel 'Rigor mortis' consists of 119 sentences, 2156 words (tokens) , the average length of sentences is equal to 18.1 tokens and the average word length is 4.6 characters.

B. *The novel 'Sono stata brava'*

The second studied novel is "Sono stata brava".It is a romantic novel written by 5 girls. This book consists of five alternative narrative plots.

As in previously analysed novel, the most voted sections have been written bypeople that are very popular from the social point of view or within the 20lines platform or on Facebook or on theirpersonal blogs in the network.

In this case weanalysedhowthe community determined its choices:the whole book with all possible plot contributions has been downloaded from the platform. This novel was composed by 75 sections distributed as follows:

- Level 1: incipit
- Level 2 -> 54 subsequent plots
- Level 3 -> 7 subsequent plots
- Level 4 -> 4 subsequent plots
- Level 5 -> 4 subsequent plots
- Level 6 -> 5 subsequent plots

The second level has been rewritten in 54 different plots;our analysis showed that only sixteen of them were semantically related to the incipit.The subsequent plot chosen by the community was not one of them. The success of the second level plotwas determined, in our opinion, by the large social popularity of the writer that on20lines has 1297 followers.

IV. Results Discussion

A. *'Rigor mortis'*

The richness of the vocabulary of the novel 'Rigor mortis' has been analysed .Lemmas belonging to the basic vocabulary (VdB) are evaluateas 65.2% , the entries related to the VdB compared to the repertoire of Fundamental Vocabulary resulted 80.9 % , of Widely used Vocabularyand of High-availability Vocabulary areequal to 14.3% and 4.8 % respectively.

The lexical density, defined as the ratio between full words and functional words, is 0.571 showing a high load of information expressed by text.

The analysed novel shape is a first-person narrative of the journey. It is practically a transcript of the conversations that took place. Usually the lexical density in daily speech is 0.3-0.4 and therefore, the calculated density of about 0.6 underlines that the reader is subjected to ahigher cognitive load.

Table 1 Co-occurrences of the word Rocco (the teacher name)

context	word	context
stavo andando al funerale di	Rocco	di nome e di fatto
per il funerale di	Rocco	Claudia, Livia, Alessia, Giulia, Mari
non c'è da stupirsi se	Rocco	di nome e di fatto
qui nelle langhe il nome	Rocco	ricorda molto la parola sasso
dico, perché adesso sono più	Rocco	perché morto, freddo e duro
traditori e vermi? Cos'era	Rocco	il mio ex professore di
molto, molto triste alla fine	Rocco	ha fatto anche del bene
quando abbiamo scoperto che	Rocco	aveva un piede in più
doveroso andare al funeraledi	Rocco	e ringraziarlo di essere stato
Cuneo, il paese natale di	Rocco	il funerale dice Alessia
di quel sexy bastardo di	Rocco	no basta concentriamoci sulla pubblicità
fisico come una pornostar	Rocco	accarezzo gli addominali fotografati e
soprattutto photoshoppati del modello	Rocco	li aveva ben definiti proprio
qualcos'altro, alla faccia di	Rocco	lo gnoccone! Siamo in ritardo
questo avrebbe più peso	Rocco	facci una grazia, solo una

In the list of the most frequent words, the first seven elements are already explanatory of the plot: they are related to the characters at the funeral and the trip in the car.

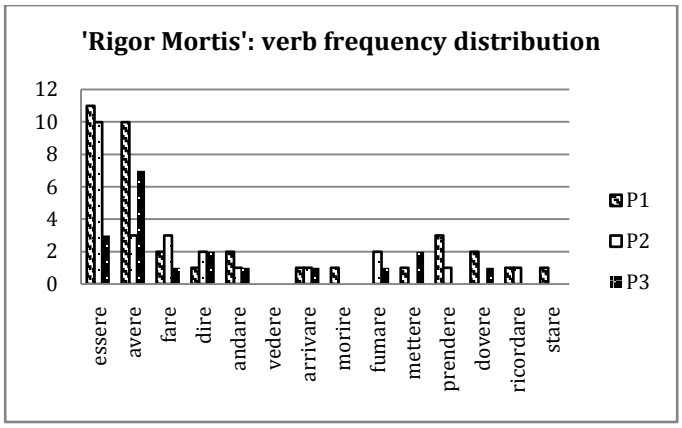


Figure 4 Verb frequency distribution for Levels of plot 1 (P1),2 (P2) and 3 (P3)

The graphs (fig.1 and fig. 2) represent the distribution of the frequencies of the most common nouns used in different levels of the text written by six different authors. These words are: 5 first names: Rocco, Alessia, Claudia. Livia, Giulia ; 2 city names: Roma and Cuneo; and funerale/funeral, macchina/car, nome/name, cazzo/fuck, faccia/face, ora/ now, ragazzo/boy, volta/time, birra/beer, esame/exam, fine/end, modello/model, professore/professor, stronzo/asshole and Faus is a dialect term that means lying. From the graphs we can see that the words "nome", "fatto" and "Faus" are used only in level 6, the words "birra" and "stronzo" only in level 4 and the use of the other words is fairly uniform in all levels of the text.

Similarly, from table 1 it can be observed that by analysing the left and right vicinity of the most frequent word, namely Rocco, the name of the teacher, it is possible to characterize

the plot.

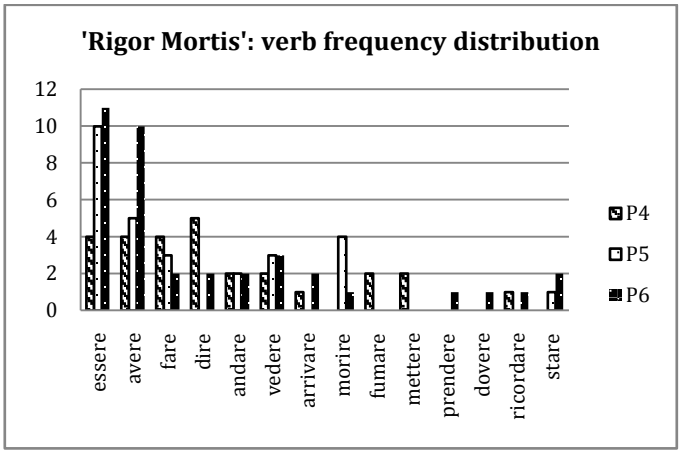


Figure 5 Verb frequency distribution for Levels of plot 4 (P4), 5 (P5) and 6 (P6)

Fig. 4 and fig. 5 plot frequency distribution of the most frequent verbs. They are: essere/to be, avere/to have, fare/to make, dire/to say, andare/to go, vedere/to see, arrivare/to arrive, morire/to die, fumare/to smoke, mettere/to put, prendere/to take, dovere/ to have to, ricordare/to remember, stare/to stay. These plots show that the number of different verbs used in 6 levels, are respectively 12, 9, 9, 10, 7 and 12, and that the most frequent verbs are to be and to have, followed by to do and to say. It is also possible to observe that the only verbs present in all the different parts are: to be, to have, to do and to go.

In table 2 the parameters concerning the morpho-syntactic categories are shown.

Table 2 Measure of the morpho-synctactic categories

'Rigor Mortis': the morpho-syntactic categories	
Nouns	16,0%
Own Names	5,9%
Adjective	5,2%
Verbs	15,7%

B. 'Sono stata brava'

Also "Sono stata Brava" was subjected tothe linguisticstatistical analysis . This novel has a total 67 periods , 1131 words (tokens) , an average length of periods in token of 16.9 and an average word length in characters 4.3 .

Table 3Measure of the morpho-synctactic categories

'Sono stata Brava': the morpho-syntactic categories	
Nouns	14,4%
Own Names	0,3%
Adjective	7,2%
Verbs	21,1%

Lemmas belonging to the basic vocabulary (BSI) are equal to 87.4 % of the total. The repertoire of Fundamental Vocabulary resulted 82.0 % , of Widely used Vocabulary and of High-availability Vocabulary are equal to 12.1% and 5.9 % respectively.

The parameters concerning the morpho-syntacticcategories have been computed and are shown in table 3.

Table 4 Co-occurrences of the word Brava

context	word	context
stata	Brava	Ho fatto la lavatrice, sono
Senza di me, sono stata	Brava	Ho rifatto il letto e
In quel piazzale, sono stata	Brava	, fin da piccola mia madre
Non lo facevo, sono stata	Brava	, sono tutti orgogliosi di me
Orgogliosi di me, sono così	Brava	Che vorrei sparire. Sono stata
Vorrei sparire. sono stata	Brava	Sono stata brava in tutto
Sono stata brava, sono stata	Brava	In tutto con gli amici
Gli amori. Sono stata così	Brava	Da darmi completamente. Sono stata
Sa darmi completamente. Sono stata	Brava	, senza maschere, senza bugie.
Senza riserve e sono stata	Brava	Ho riso e pianto senza
Mai nascondermi e sono stata	Brava	Sono stata talmente brava che
sono stata brava. Sono stata talmente	Brava	Che vorrei esserlo stata
Credo di non essere stata	Brava	In gamba o precisa
Polpastrello mi dica sei stata	Brava	Anche solo per un bacio
Della giornata. Sono sempre stata	Brava	Anche a scuola amavo sentirmelo
E così di sei stata	Brava	Ho incominciato a riceverne di
C'è più. Sono stata	Brava	Tutte le volte in cui
Io non sono stata molto	Brava	A sistemare le cose.

The display of the co-occurrences of the word Brava, the most frequent word, underlines the completeness of the plot due to the fact that each node story is self sufficient.

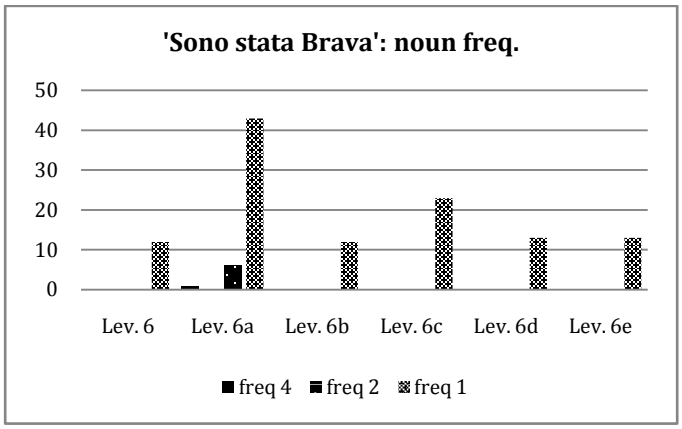


Figure 6 Noun frequency distribution for six ending plots where freq 4, freq 2 and freq 1 stand for the number of lemmas with frequency 4,2,1 respectively

We analysed in detail the 6 end plots (namely 6, 6a, 6b, 6c, 6d, 6e) from the morphological point of view. Where lev 6 is the plot chosen by the social web.

The analysis of six different end texts (levels6) is plotted in in figures6,7,8 and 9. These pictures show that what has been decreed as the "winner", (the Lev 6 text), is less rich than others from the lexical point of view, for the grammatical

categories of nouns, verbs, adjectives and adverbs. The richest text seems to be Lev 6a followed by Lev 6c.

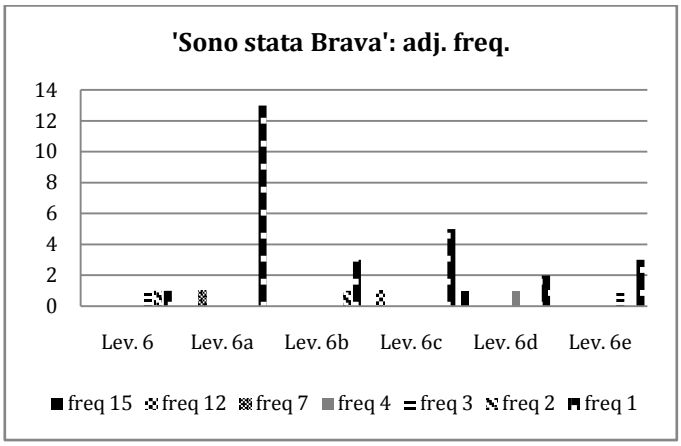


Figure 7 Adjective frequency distributions for six ending plots where freq 15, freq 12, freq 7, freq 4 freq 3, freq 2 and freq 1 stand for the number of lemmas with frequency 15,12,7,4,3,2,1 respectively

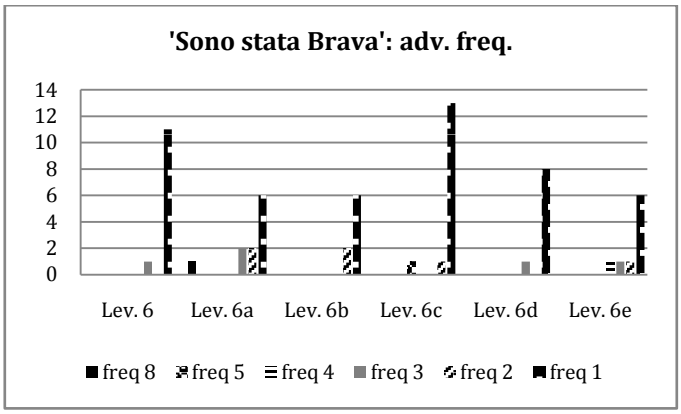


Figure 8 Adverbs frequency distributions for six ending plots where freq 8, freq 5, freq 4, freq 3, freq 2 and freq 1 stand for the number of lemmas with frequency 8,5,4,3,2,1 respectively

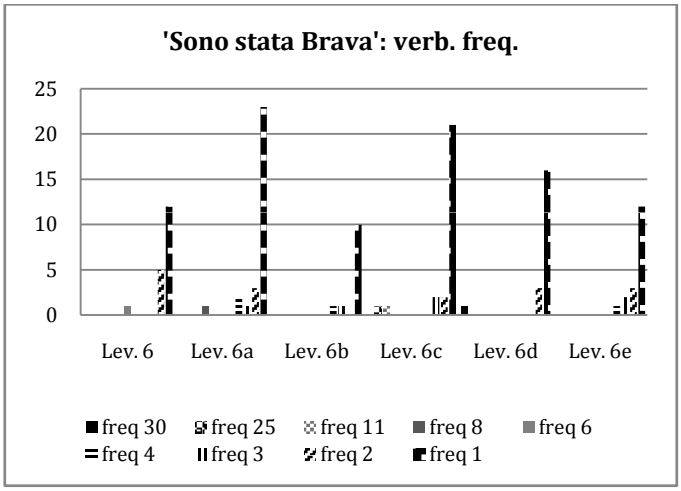


Figure 9 Verbs frequency distributions for six ending plots where freq 30, freq 25, freq 11, freq 8, freq 6, freq 4, freq 3, freq 2 and freq 1 stand for the number of lemmas with frequency 30,25,11,8,6,4,3,2,1 respectively

Graphsof figures 10,11,12,13 show the morpho-syntactic analysis of 6 winning plots of "Sono stata Brava". It can be noticed that thelexical richness of the 6 levels has a non-uniform distribution referred to the grammatical categories (nouns, verbs, adjectives and adverbs). In particular level 3 and the incipit (Lev1) of the plot turn out to be richer than the others.

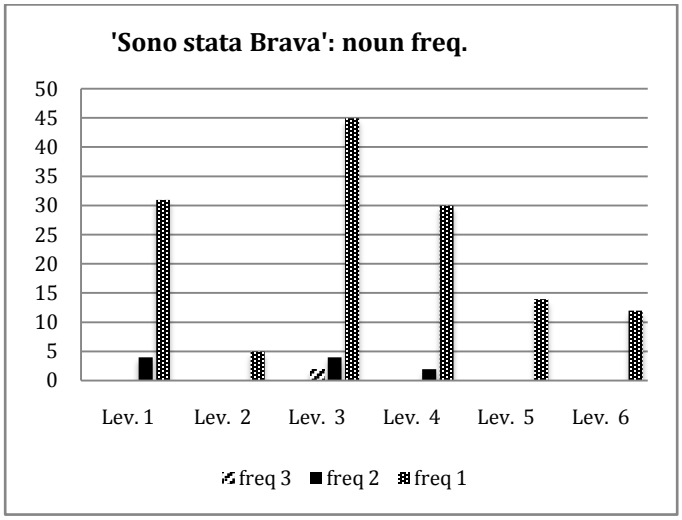


Figure 10Noun frequency distribution for the final novel where freq 3, freq 2 and freq 1 stand for the number of lemmas with frequency 3,2,1 respectively

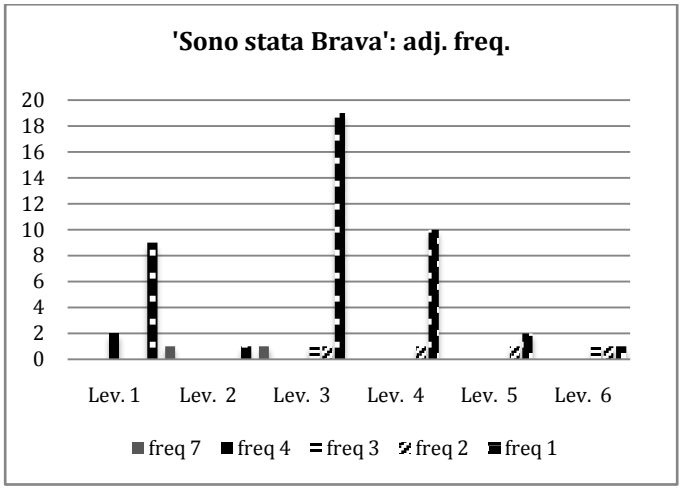


Figure 11 Adjective frequency distributions for the final novel where freq 7, freq 4, freq 3, freq 2 and freq 1 stand for the number of lemmas with frequency 7,4,3,2,1 respectively

The whole final novel, with its all 6 narrative plots is not linguistically complete, i.e. the text of each node does not provide the full set of information needed to understand the novel focus and to achieve a given goal.It does not show adequate thematic, semantic and logic consistency.

This novel looks more as a kind of collaborative hypertext novel, where the analysis of the characters, the psychological introspection and especially the number of points of view involved are quite different in every level. Each level tells an episode of the novel by itself with a self-contained plot with its own narrative value. The final story is, with its complexity, an aggregation of inconsistent stories but still able to give the reader the opportunity to find its most suitable plot. Each

ISBN: 978-1-61804-236-1

subsequent plot seems to belong to the stories of the third kind defined by Andrea Smorti[22].

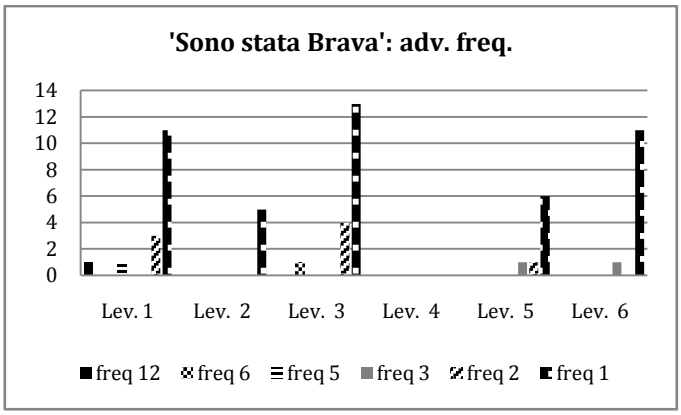


Figure 12 Adverbs frequency distributions for the final novel where freq 12, freq 6, freq 5, freq 3, freq 2 and freq 1 stand for the number of lemmas with frequency 12,6,5,3,2,1 respectively

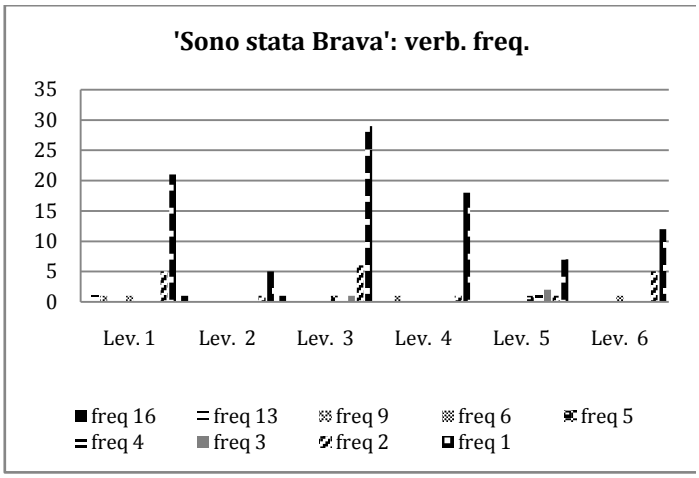


Figure 13 Verbs frequency distributions for the final novel where freq 16, freq 13, freq 9, freq 6,freq 5, freq 4, freq 3, freq 2 and freq 1 stand for the number of lemmas with frequency 16,13,9,6,5,4,3,2,1 respectively

The stories of the Third Kind have inconsistent plot, i.e. they show a problem but they do not give its solution; therefore they do not have a significance that is coherent. This type of plots are built in such a way that the reader / listener have to find its own solution.

The novel 'Sono stata brava' could be associated with the "possible worlds" of Goodman[23]. According to this theory on narrative thinking, people are able to build their own effective and fair models of the world, starting from a world that is given. By adopting different points of view, each person creates different templates of the same world: all of them aim to explain it and are very good. These templates are the result of a 'production' of worlds, through various ways.

V. Conclusion

In this short research we showed that the characteristics of the analysis of the textual narrative plots may be used for studying collaborative narrative texts when they are built as stories connected by logical consistency and they are capable of expressing a content unified and understandable.

In particular, the statistical analysis of language is a tool suitable for the analysis of the new 2.0 collaborative writing.

The quantitative and qualitative analysis of language is a useful tool for the study of a new collaborative writing 2.0, in fact the study of the frequency of certain full words and their concordances allowed the identification of the keywords of the content of the narrative text (i.e. 'Rocco'). This preliminary study opens up new ideas for research, such as the comparison of narrative texts written by the same author in the same book in both 20line.com and in the Relay platform, or the analysis on how the social component can affect the narrative text.

References

[1] Lowry P.B. Curtis A. and Lowry M.R. Building a Taxonomy and Nomenclature of Collaborative Writing to Improve Interdisciplinary Research and Practice *Journal of Business Communication* 2004; 41; 66 DOI: 10.1177/0021943603259363.

[2] McGoldrick Nikoo; James A. McGoldrick (May 2000). *Marriage of minds: collaborative fiction writing*. Heinemann. ISBN 978-0-325-00232-3. Retrieved 19 September 2011.

[3] Ritchie, Stephen M; Rigano, Donna L (2007). "Writing together metaphorically and bodily side-by-side: an inquiry into collaborative academic writing.". *Reflective Practice***8** (1): 123–135. Retrieved 28 February 2013.

[4] Calvino Italo *Lezioni americane. Sei proposte per il prossimo millennio*, Milano, Garzanti, 1988.

[5] Borges Jorges Louis: "*El Jardín de senderos que se bifurcan"* Editorial Sur, 1941, argentina.

[6] Joyce Michael "*Afternoon, a story*"1987 published by Eastgate Systems in 1990.

[7] Craig McLean, "It's A Funny Old Game: Footballer LUTHER BLISSETT has become a hero to Italian anarchists. Why?", Word Magazine, April 2003.

[8] Deseriis Marco,"'Lots of Money Because I am Many:' The Luther Blissett Project and the Multiple-Use Name Strategy". In *Cultural Activism: Practices, Dilemmas and Possibilities*, edited by Begum O. Firat and Aylin Kuryel, Amsterdam: Rodopi, 2010. p.65-94

[9] http://en.wikipedia.org/wiki/Luther_Blissett_Project

[10] **http://en.wikipedia.org/wiki/Wu_Ming**

[11] Gregorio Magini and Vanni Santoni, "Il metodo SIC", http://www.scritturacollettiva.org/documentazione/metodo-sic

[12] Pispisa Guglielmo'Gli altri, gli stessi. L'identità al tempo dell'autore collettivo" Mantichora http://ww2.unime.it/mantichorapg. 557, 1, 2011, ISSN 2240-5380

[13] www.passalapenna.it

[14] http://it.20lines.com

[15] Propp, Vladimir. Morphology of the Folktale: Revised and Edited with Preface by Louis A. Wagner, Introduction by Alan Dundes. Vol. 9. University of Texas Press, 2010.

[16] Barthes Greimas Todorov Eco Genette Gritti L'analisi Del Racconto Bompiani 1969Russell&Ryan *2010* "Analyzing Qualitative Data: Systematic Approaches", Los Angeles

[17] Russell&Ryan *2010* "Analyzing Qualitative Data: Systematic Approaches", Los Angeles

[18] "LinguA: Linguistic Annotation pipeline", Italian Natural Language Processing Lab, http://www.italianlp.it/?page_id=1029

[19] Dell'Orletta F. "Ensemble system for Part-of-Speech tagging". In: Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009 (Reggio Emilia, Italy, December 2009)

[20] Attardi G., Dell'Orletta F. "Reverse Revision and Linear Tree Combination for Dependency Parsing". In: NAACL-HLT 2009 – North American Chapter of the Association for Computational Linguistics – Human Language Technologies (Boulder, Colorado, June 2009). Proceedings, pp. 261 – 264. Ass. for Computational Linguistics, 2009

[21] De Mauro, T.*Vocabolario di Base della lingua italiana*, 1989

[22] Smorti A. Il pensiero narrativo Giunti editore , Bologna 1998

[23] Goodman, N. "Ways of Worldmaking", ISBN:9780915144518

Distributed sensor network – data stream mining and architecture

T. Lojka, I. Zolotova

***Abstract*—**Data mining techniques are traditionally centralized and need huge computation power. Centralized data mining computation can be spread out across area and become a distributed network of small computation resources, for example sensors. In this paper, we deal with a distributed sensor network in which we implemented a data mining technique to reduce the network overload. The network we designed consists of sensors with small computation resources and an implemented data mining algorithm. Data mining is used to find valuable information in an unfinished flow of sensor data. Groups of sensors are connected to gateways. The gateways create Wi-Fi connections to the central unit of our network. For communication between a gateway and a central unit, we use the Service Orientated Architecture (SOA) with Simple Object Access Protocol (SOAP). This solution has been tested for use in monitoring patients.

***Keywords*—**Data mining, distributed network of sensors, monitoring system, Service Orientated Architecture (SOA).

I. Introduction

NOWDAYS we live in the world of complex monitoring system with many sensors around us. A lot of places, buildings or machines have own network of sensors. These sensors get a part of our everyday life. Actually we used them to immediately find, diagnose and solve problems. For example in health care are used sensors to monitor patient health state. Or we use sensors to identify traffic jams, to find the most effective way to work or home, finding incidents in workflow and so on.

Monitoring system is used to monitor environments or objects in environment through sensors and help people to be more effective in their work. Exactly the sensors are for assurance that unawares anomaly will be immediately noticed even predicted with the monitoring system. To identify or predict anomaly we need to have a real-time monitoring system. The system should monitor problems of itself (system resources) and monitored environment or objects [1]. Because of complexity of monitoring systems it usually consists of large number of sensors. The sensors produce huge volume of data. This data are from distributed sensors networks. Therefore sensor data might have a different consistency and they are dynamic [2].

Monitoring system contains sensors which are used for monitoring environment and its instances. Amount of data which produce sensors in monitoring systems is very big. It can be said that amount of data is enlarging with every day. Data from sensors are usually send directly in data streams to system which processes, uses and saves them. The overload of network is bigger and some not important data are sending over the whole network from sensors to master or from peer to peer. This problem can be explained with phrase "data rich but information poor" [3].

Bigger overload of network increases a time-latency of whole system. This time-latency is dangerous for real-time control and handling unawares errors, alarms and events might be unsatisfactory.

Geographically distributed and heterogeneous system produces huge volume of data which causes communication problems. This problem can be solved with reducing communication overload (II).

Monitoring system creates unique research opportunity to solve problems how to design network of sensors witch will solve communication, lower power consumption, hazardous deployment, sensors failure tolerance, distribution of sensors and processing data from sensors.

II. Distributed and centralized sensor data processing

Researches in sensor network are trying to decrease number of transmitted messages, retransmitted messages, latency, but increase throughput of network. Today we can find concepts and solutions that describe centralized or distributed data processing from sensors.

Centralized data processing is implemented in monitoring system that has one centralized computing resource. This centralized computing resource collects data from all sensors in the network and process them. After processing them the centralized computing resource saves them to DB or a computing unit might sent fused data to the clients, which request the data.

Related work is Centralized Dynamic Clustering (Bajaber, F., Awan, I.) [4]. Nodes are group to cluster. Cluster is managed by cluster head. Cluster head is responsible to collect data from all nodes inside cluster [4].

Reference [5] cares about distributed sensors network like a

This paper was supported by grants KEGA-021TUKE -4/2012 (50%) and VEGA - 1/0286/11 (50%).

T. Lojka is with the Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Košice Slovakia (email: tomas.lojka@tuke.sk).

I. Zolotova is with the Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Košice Slovakia (email: iveta.zolotova@tuke.sk).

network that consists of autonomous sensors. These autonomous sensors are placed over a large area [5]. Each node of the network has a computing resource, but the resources are limited. Opposite, a sensor measurement task is performed by large number of tiny computing resources (sensors) [5]. Benefits for this autonomy sensor network are:

- Overall monitoring system is more robust to failures [5].
- Sensors are small, have small computing power, but have less power consumption. It is easier to replace them and can be used in small spaces [5].
- Overload of sensor network traffic is spread out [5].

Difference between centralized distributed processing is described in table I. [2].

TABLE I.
DIFFERENCE BETWEEN CENTRALIZED AND DISTRIBUTED SENSOR DATA PROCESSING

Attribute	Centralized	Distributed
Energy resource	Non-bounded	Bounded
Computation power	High	Low
Flow of data	Stationary	Continues
Data length	Known	Unknown
Update speed	Low	High
Passes	Multipass	Single
Time of processing	Non-real-time	Real-time
Memory resources	Large	Small

Distributed sensors data processing can be process explicitly (each node process its own data) or implicitly (nodes are connected and share resources). More used is explicitly algorithm, but disadvantage opposite the implicitly algorithm is that explicitly algorithm do not have better information result in describing monitored environment.
Explicitly algorithm does not cooperate with surrounded nodes. Therefore explicitly algorithm can't reach better monitoring results, but it decrease traffic overload.

Disadvantage of explicitly algorithm is in cooperation, fault tolerance and quality of measurements [4].

In our work we focus on sensors which have different physical principles and do not monitor the same instance.

III. DECREASING AMOUNT OF SENZOR DATA WHITHOUT LOSSING INFORMATION

A. *Methods for decreasing amount of sensor data*

Big amount of data which has no information value for control or prediction of system is wasting of resources and bring into system bigger uncertainty and variety (Heisenberg uncertainty principle). Options how to increase amount of information and decrease amount of useless data are:

1. Finding the best sampling frequency. Sensor will send measured data only if it is needed. This will decrease amount data, but do not decrease information value of data.
2. Using threshold to find only interesting data. Unnecessary information is cut with low threshold value. There might be used mean shift vector [6]:

$$M(\mu) = 1/n_x \sum_{i=1}^{n_x} (x_i - \mu_0) \quad (1).$$

Where μ_0 is initial value, n_x is number of measured values with definite sampling frequency, x_i is measured value.
3. Sensors immediately process data and only information is sending from sensors. Sensors has implemented algorithm that gives output with description that is useful for actuators.
4. Sensor will send packages of measured data in compressed form. This means that only change of information is noticed.
5. Events based solution for transmitting data from sensors. Every change of information in sensors is presented like an event. This event calls master for listening and sends valuable information to the master network.

   The event reduce amount of data, but not decrease

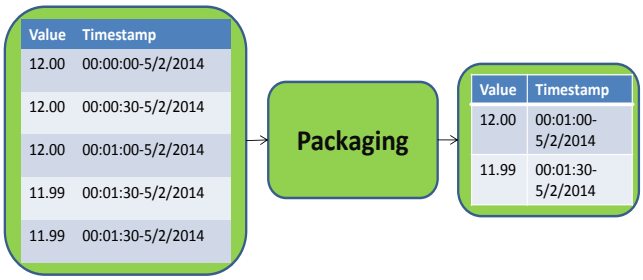


Fig. 1 Packaging data sending from sensors

   an information value.
6. Using data mining to reduce amount of data coming from sensors [4]. This option leads to reduce amount of information that are transferred from sensors, decrease overload of network and decrease a variety and uncertainty of whole system.

B. *Comparison of methods for decreasing amount of sensor data*

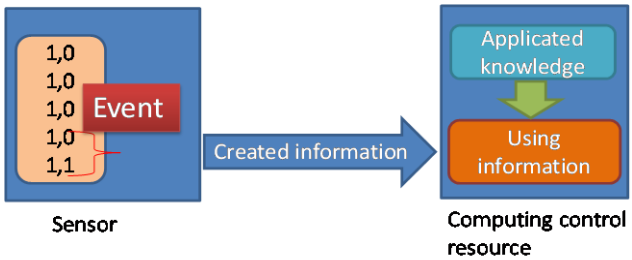


Fig. 2 Event based solution for transmitting data

In this paper we focus on reducing amount of data, finding and sending only valuable information and so reduce network traffic. We compare positive and limitations of mentioned

options to reduce data traffic from sensors. We chose data mining in combination with time sampling, threshold and events.

TABLE II.
OPTIONS FOR REDUCING DATA TRAFFIC FROM SENSORS

Opt.	Appearance		
	Description	Positive	Limitations
1.	Uses time sampling	Reduce amount of data	Involve data with not needed information
2	Uses threshold	Less data	Not all data has valuable information
3.	Produce information that the actuators can use.	Produce control information	Needs bigger computing resources
4.	Uses reducing of data by monitoring	Reduce amount of data	Not all data has valuable information
5.	Uses event to send information	Work with information	Need higher computing resources
6.	Uses data mining	Work with information	Need higher computing resources

IV. DATA MINING IN SENSORS

Sensors produce data. This data are processed to reach information. Then the information is used to reach knowledge. The knowledge says something about monitored environment or instance. This process is known as data mining. Data mining can be used for:

- Describe the data from sensors
- Predict data from sensor, exactly predict sensor output.

Data mining is process which is not conventionally design for sensors. Usually data mining works with static data types, has unlimited processing time, high computational power, stationary data flow, non-real-time response time, and has mulitipass algorithms [2]. These features of data mining are not feasible with sensors. Sensors has constrains like power, computing resource, memory, and communication [2].

Sensors quickly produce queues of data. This data queues should be real-time processed by data mining techniques. Hence, cyclic or more step data mining algorithm are inconvenient.

Sensors are usually representing distributed system. In distributed system are data changes very fast. To create distributed network of sensors with good time response, we need to implement an online data mining. Despite, sensors computation power is limited (less like a central computer of sensors networks), sensors create distributed system of computation power.

V. DATA STREAM MINING TECHNIQUES IN SENSORS

Data mining in sensors has big potential. It can be used for monitoring, classifying, power saving, communication reduction, and prediction [2]. Using data mining in sensors creates distributed network of computing power and solve problem with growing amount of data, which is problem to store and index. In our solution we consider on communication reduction in distributed network of sensors.

Network contains sensors which dispose with low memory and do not have database system inside them. Therefore we should think about stream of data and low memory sources. This memory source represents a buffer with fixed length. The buffer type is FIFO and contains last n measured values. For this goal we use sampling to estimate entropy a_n.

$$S = \langle a_1, a_2, a_3 ... \rangle \rightarrow S_i = \{a_1, a_2, ..., a_n\},\ n = \{1,2,...\} \quad (2)$$

Every time the buffed is full of values is consider as a frame of length n.

We identify process of data handling to reduce communication workload which is depicted on Fig. 3. Our process consists of points which are written below.

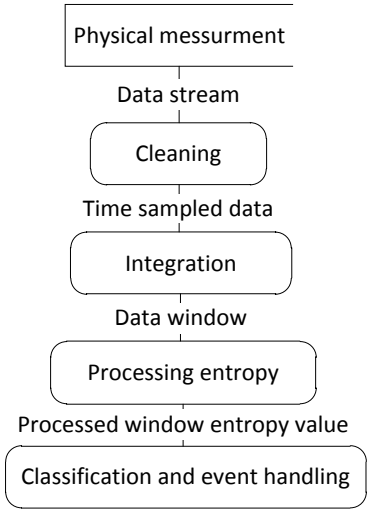


Fig. 3 Process of data handling to reduce communication workload

Physical measurement – presents a process of measuring external environment based on physical principles. Output from this process is created data stream.

Cleaning – presents a process of cleaning irrelevant data from data stream. Sensors produce a stream of data. Irrelevant data are cleaned by setting propriety sampling frequency. With data sampling can be reduce usage of costly computing and memory.

Regular next step is **data integration**. Because our explicitly definition of sensor network does not produce heterogeneous data we partially skip this step in this work. We integrate data only into frame. Frame represents specified number of measured values ordered into array.

Data transformation – presents a process of transformation data in suitable form for data mining. In our condition it is represent by rounding measured values to defined decimal place, which will speed up computing and decrease memory consumption.

Processing entropy – we choose algorithm for selecting data, which are important for analyzing frames information value.

We need to separate data with low information value from

data with higher information value. One option was to choose, Shannon entropy non-conditional empirical entropy (3).

$$H(S_i) = -\sum_{j=1}^{n} p_j \log p_j \quad i, j = \{1,2,3,...\} \tag{3}$$

Where n represent count of measured values in one frame and p_j is probability of value in current frame. Next statement represents frequency of *i*-th value in frame. The empirical probability distribution is defined in (4).

$$p_i = \frac{f_i}{\sum_{j=1}^{n} a_j} \quad , j = 1,..., N \tag{4}$$

Another option was to choose entropy defined by Tsallis (1988) [5].

$$S_\alpha(p) = 1/(1-\alpha)\left(\sum_{j=1}^{n} p_j^\alpha - 1\right),\ 0 \le \alpha,\ \alpha \ne 0 \tag{6}$$

Tsallis defined entropy which is additive and contains parameter α. This parameter enable to setup influence of probability distribution [7].

Because of fluent of data stream we chose one pass algorithm. This way we can save sensor computing power for next frame of streamed data. One pass algorithm decrease time computing and enable sensor to work in fast real-time (Measured delay was 1.5ms opposite each measured value).

Classification and event handling – present a process of evaluation of truly valuable information and representation. In this process we can compare already computed entropies to find which is over defined dash. If the frame has entropy which is higher or equal that (static/dynamic) defined dash, then the sensors create an event. Other words sensor has recognized unusual behavior in measured data stream. Using this event, the sensor tells about data stream anomaly to the master of sensor network or central PC. If the entropy has less value as defined dash the whole frame is erased.

Creating event will reduce overload of network traffic. Events are created only if valuable information will occur in sensor and every data sample will be send to central network computer.

Fact is, that lot of data, which might be useful in the future are erased. In this solution we propose network which is not for critical measurement and control in real-time yet.

VI. Sensor Data Mining Architecture

We defined architecture which consists of sensors, gateways and central computer (Fig. 4). All nodes in architecture have some computation resources. Therefore it creates powerful network for data mining consisting of sensors with small computation power and memory.

A. Sensors

Sensor is used for monitoring patients. In our solution we have already implemented accelerometer sensor. Accelerometer measures vibrations, which are caused by movement of patient (for example we used accelerometer to identify patient movements during the patient was standing up from bed).

First of all we will describe sensor architecture (Fig. 4). Sensor measures vibrations. We propose sensor architecture to process data stream of measured vibrations. Whole data processing in sensor is described on Fig. 5.

Sensor measure vibrations and produce data stream of measured values. Then they are processed by computing unit. In the computing unit, the measured vibration data stream is sampled. Data stream is sampled to reduce amount of data. We need to reduce the amount data to save energy and computation resources. Sensor measurement dynamic and

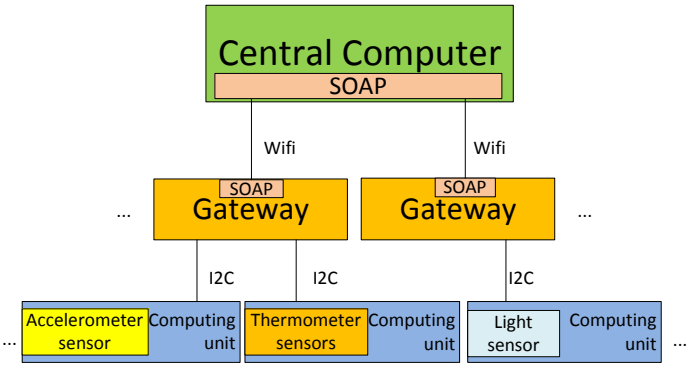


Fig. 4 Sensors network architecture

amount of wanted information allow us to sample input data stream.

Sampled data goes to the next part of computing unit (finding information). In finding information part of processing unit we implemented data mining algorithm to find valuable information in stream. Other data from stream are removed to free sensor computing unit memory. Finding information part of computing unit is based on entropy. We used Tsallis's definition of entropy. Detail algorithm is described in chapter VIII.

Process of data filtering is followed by process of classifying information. Information is classified into groups. Every group creates events with defined priority level. Detail algorithm is described in chapter VIII.

Outputs from computing unit are events. Special parts of sensors are Conditions and Working storage. Sensors usually consist of unit for monitoring sensor system and unit for controlling system [8]. In our sensor architecture (Fig. 5) is monitoring system represents by (Stream processing, Event, Working storage) and was described above. Unit for controlling system is depicted with thicker border on Fig. 5. Controlling unit is presented by Conditions block. It is static block of sensor data which is used to initialized, setup, storing data mining rules, and defines calibrations attributes.

Working storage is used for storing working data. During communication are important results saved to Working storage. Then they are used in process of finding information or classifying information [1].

Process of classification has an information input and

event output. This unit defined the importance of information and classifies it into defined event classes. We have already defined only two event classes:

1. Alarm class – this class represent information which contains critical data. This group is defined by rules which are stored in sensor Condition block.
2. General class – represent a class where information from sensor does not represent any critical situation.

Rules for classification was set experimentally and saved in Condition block in sensor. Rules can be edited, because Condition block presents a unit for sensor controlling system.

B. *Gateways*

Gateway represents communication node, which connects network of sensors and Wi-Fi network. Gateway is part of monitoring distributed sensor network which integrates:

• Communication part for connecting two different networks [11].

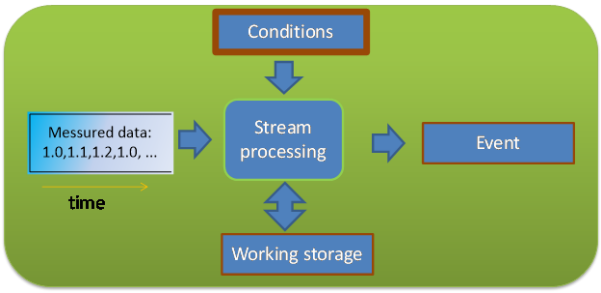


Fig. 5 Sensors architecture

• Logic part represents implicitly sensor data processing [11].
• Hardware part creates circuit for connecting sensor and Wi-Fi adapter, computation resource and memory [11].

In our solution we used SOA to present communication with central PC. Service is running on central PC. Clients (gateways) send requests to service. Service accepts requests process them and return an acceptation message to gateway.

The best way how to send message in our option was to use handshake. Therefore gateway waits for handshake of central computer [11]. If gateway will receive handshake then the message will be erased from queue waiting for message. After the handshake will not come, the message will be send again. Handshake solves network errors and guaranties that message about event will be delivered and accepted by central PC [5].

C. *Central PC*

Central PC is PC that collects data form whole network and process them. Central PC hosts a service (SOA) and has integration potential in satisfying communication with other central hosts or other systems.

Central PC might be also presented in cloud and information will be process, and stored inside cloud (Fig. 6). Cloud has big computing power to do data mining with sensor data, but our purpose was to reduce communication and sends only messages with important information. This should solve communication bottleneck. Therefore this implementation reduce commination overload and create architecture for a real-time sensors data processing. This proposed architecture may contain more sensors and gateways per one central PC due communication overload reduction.

VII. Sensor Data Mining Implementation

In our solution we implemented Tsallis entropy to compute changes of information in data stream. Because Tsallis entropy operates with stream of defined length we create a frame. Our frame contains defined number of measured values. Every new

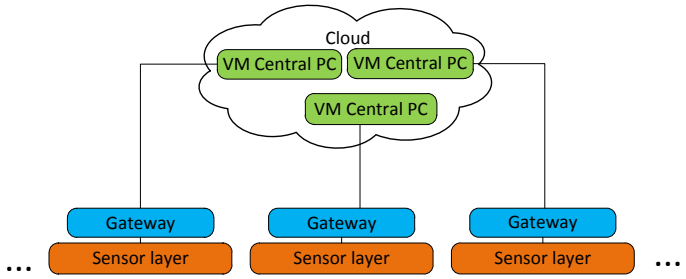


Fig. 6 Proposed sensor architecture implemented in cloud

value or values is/are written in array and the last value/values is/are removed. Range how the sensor will computes with new values and removed old is defined inside computing algorithm and it is configurable due configuration system of sensor.

After new values in frame are written, the computing unit starts to compute entropy. If computed entropy is less, there is no big change of information in selected stream. To be more precisely we thing about options how to build frame of data and how to not overview important information in data stream.

Our solution is based on floating frame with fields. In the configuration unit is set how values should be inserted into frame and how values should be removed from the frame. With this definition we defined a sensitivity of information change in stream.

Another part of sensitivity is controlling if the frame reach upper dash for creating event. This is used for classification, if selected value will create an event. If the computed entropy is upper than a defined dash, the sensor computing unit will start to classify data in the stream to create appropriate type of event.

Tsallis definition of entropy allows setting parameter α. This parameter was experimentally defined to value "0.9". Using the reference [8] we choose Tsallis entropy to calibrate sensitivity of computing. If the parameter α is positive then sensor will indicate values that occur often. If the α if positive then the sensor will catch measured values that happen seldom [8].

VIII. Sensor Data Stream Mining Results

We implemented Tsallis entropy which we apply on data stream from sensor. Algorithm for computing entropy was implemented directly in the environment of sensor. First we reduce amount of information.

We definite sensitivity of frame to two and frame length to five. Next we identify entropy dash, to filter low changes of entropy caused by noise.

If the value of entropy is upper than defined dash we used weighted average of current processing frame to find the main important value of frame.

In the Fig. 7 are presented our results. We measure only x axis of accelerometer. The value was measured during patient standing up from bed.

We log sensor values and write them to the file. The values written in logs we used for later analyzing of tested algorithm. Every value has time-stamp and measurement voltage value from accelerometer x axis.

We used only x axis to test functionality of algorithm. After measurement we plot measured values from logs and compare it with values from data process with tested algorithm and computed entropy (Fig. 7). We can calibrate sensor to be more sensitive or less sensitive, to indicate more seldom or often occurred measured values, or set rules for classification. With this calibration we can tune amount of data which will be sent over network and also tune an information value of sending data.

We recognize time delay due the data mining computing. The time delay in average we round to 1ms. But the algorithm recognizes changes in data stream from sensor.

Our gateway functionality accepts data from sensor. After gateway receives value, the gateway will send it to central PC and wait for acceptation from sensor. Central PC functionality has not been implemented yet.

For our testing we used components from smartphone, which has accelerometer, processing unit, memory and Wi-Fi. In the future we would like to implement this solution on, sensor boards (microcontroller-based sensor board with set of sensors and communication interfaces) [11] .

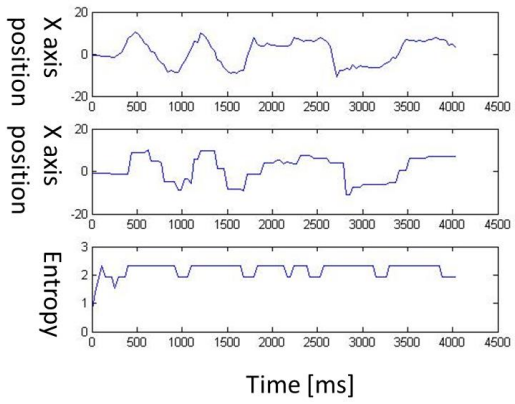


Fig. 7 Deviation comparison of real-time data from sensor, reduced data communication in real-time and entropy.

IX. Conclusion

Our purpose is to find way how to reduce communication overload in distributed sensor network. To reduce sensor network overload we implement data mining techniques and event based communication. With data mining we are trying to find only valuable information in sensor measured data and erase data with no information value. We design architecture of network where the algorithm might be implemented.

In the future we would like to test data mining techniques and find option how used this techniques over real-time data stream, tune data mining techniques for real-time deploying in hospital or home to leverage health care about patients.

Acknowledgment

This work was supported by colleagues from Laboratory of intelligent control networks and control software systems DCAI FEI TUKE.

References

[1] L. Wang; J. Zhang, Z. Zhang, G. Sun, and G. Chen, "KD monitoring system: Design, implementation, and evaluation," *Computer Research and Development (ICCRD)*, 2011 3rd International Conference on, vol.2, pp.120,124, 11-13 March 2011.

[2] A. Mahmood, K. Shi, S. Khatoon, and M. Xiao, "Data Mining Techniques for Wireless Sensor Networks: A Survey," *International Journal of Distributed Sensor Networks*, vol. 2013, Article ID 406316, 24 pages, 2013.

[3] W. J. Slotnik, and M. Orland. "Data rich but information poor." *Education Week* (2010).

[4] F. Bajaber, and I. Awan, "Centralized Dynamic Clustering for Wireless Sensor Network," *Advanced Information Networking and Applications Workshops, 2009*. WAINA '09, 26-29 May 2009, pp.193-198.

[5] J. D. McLurkin, *Algorithms for distributed sensor networks*. Diss. Department of Electrical Engineering and Computer Sciences, University of California, 1999.

[6] A. Basharat, N. Catbas, and M. Shah, "A framework for intelligent sensor network with video camera for structural health monitoring of bridges," *Pervasive Computing and Communications Workshops, 2005*. PerCom 2005 Workshops. Third IEEE International Conference 8-12 March 2005, pp.385,389.

[7] C. Tsallis, R. Mendes, and A. Plastino: "The role of constraints within generalized nonextensive statistics". *Physica* 261A (1998) pp. 534–554.

[8] T. MASZCZYK, and W. DUCH, "Comparison of Shannon, Renyi and Tsallis entropy used in decision trees". In: *Artificial Intelligence and Soft Computing–ICAISC 2008*. Springer Berlin Heidelberg, 2008. p. 643-651.

[9] A. Lall, V. Sekar, M. Ogihara, J. Xu, and H. Zhang, "Data streaming algorithms for estimating entropy of network traffic." *ACM SIGMETRICS Performance Evaluation Review*. Vol. 34. No. 1. ACM, 2006.

[10] A. Rajaraman, and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2012.

[11] A.-B. García-Hernando,J.-F. Martínez-Ortega,J.-M. López-Navarro,A. Prayati, and L. Redondo-López, *Problem Solving for Wireless Sensor Networks* (Book style), Springer, 19.12.2008, pp. 34-54.

[12] I. Zolotova, L. Lacinak, T. Lojka, "Architecture for a universal mobile communication module," *Applied Machine Intelligence and Informatics (SAMI), 2013 IEEE 11th International Symposium*, pp.61,64, Jan. 31 2013-Feb. 2 2013

Fast insight into time varying datasets with dynamic mesh

Vaclav Skala, Slavomir Petrik

***Abstract*—**Advances in computational power over the last two decades allowed fluid dynamic simulations that involve moving parts. Various mesh update methods are employed in such simulations to adapt cells around the moving parts, resulting in a separate new definition of the mesh geometry and associated values for each discrete simulation step. Common practice is to visualize every timestep of the simulation as a single static dataset. We present a novel method for interactive visualization of the evolving isosurfaces from the datasets with dynamic mesh. The effectiveness of the method is demonstrated on the real-life datasets from combustion simulation.

***Keywords* —** Iso-contour; time-varying mesh; interactive visualization.

I. INTRODUCTION

STEEP increase in computational power over the last 20 years allowed fluid dynamic simulations that involve moving parts. Examples are combustion simulations with moving piston inside an engine cylinder [7, 16], see Fig.1, or falling payload from under an aircraft wing [13]. Various mesh update methods are involved in such simulations. The goal of the mesh update procedure is to re-calculate position and shape of the cells surrounding moving parts. Resulting datasets then have separate definition of the mesh geometry and values of the simulated quantities for each discrete simulation step.

The biggest challenge in visualization of such a data is their large size due to the changing geometry of the mesh at successive timesteps. The inter-timestep correspondence of the mesh Machine cells is usually lost. Another factor that contributes heavily to the overall size of such datasets is time-varying values associated with cells.

Isosurfaces are a standard tool for the investigation of fluid dynamic datasets. Most of the existing methods for efficient isosurface extraction from time-varying datasets assume static mesh geometry over the course of simulation with time-varying values. Only a few method exist that are able to handle datasets with dynamic mesh geometry.

The method presented in this paper allows for interactive playback of the evolving isosurfaces extracted from the datasets with dynamic mesh. We do not make any assumptions about the way the simulation mesh changes between adjacent timesteps. The major contribution of our work is a novel metrics for evaluation of temporal variation in a cell's shape, which is used to re-establish lost inter-timestep cells correspondence. Those cells with similar position, geometry and value are efficiently re-used for multiple timesteps instead of being re-loaded from disk. The method proposed is particularly suitable for the engineering software, in which quick insight into investigated data is required rather than high accuracy of the computed isosurfaces.

II. RELATED WORK

The idea of dynamic simulation mesh is not new. Arbitrary Lagrangian-Eulerian (ALE) methods were developed for this purpose. Donea et al. [14] provides a good survey of the field.

Most of the existing techniques for a efficient isosurface extraction from time-varying datasets assume static mesh. In static mesh the number of cells and their geometry do not change during the course of simulation. For such datasets a family of method has been developed based on the notion of Span-Space [5]. In Span-Space each cell of the dataset is

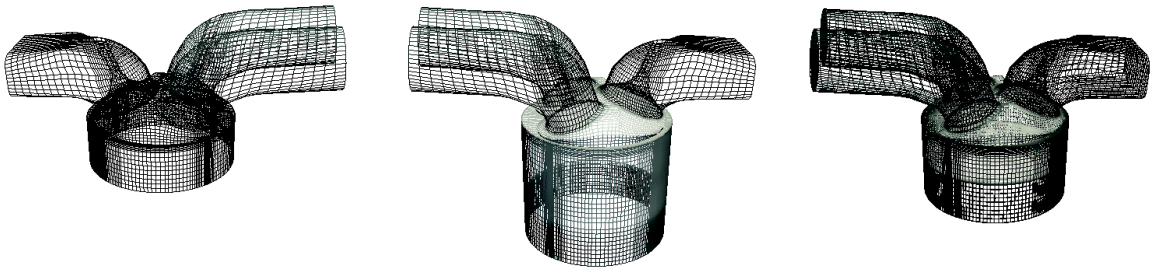


Fig.1: Dynamic mesh updated during the course of a combustion simulation. The mesh adapts total number of cells and their shape around moving parts - piston and fuel intakes.

This work was supported by MŠMT Czech Republic, projects No.LG13047 and LH12181. V,Skala and S.Petrik are with the University of West Bohemia, Faculty of Applied Sciences, Department of Computer Science and Engineering, Plzen, Czech Republic (http://www.VaclavSkala.eu)

represented as a point in plane with coordinates x, y equal to the minimum and maximum value associated with the cell's vertices. Shen et al. use lattice subdivision of the Span-Space in their ISSUE algorithm [6]. Waters et al. [19] organizes the cells into the structure of fixed-sized buckets according to their min-max value.

Originally proposed for the isosurface extraction from static datasets, the idea of Space-Space has soon appeared in the methods for time-varying datasets. The Temporal Hierarchical Index Tree (THIT) [8] assigns the cells of a simulation mesh to its nodes according to a temporal variation of their minimum and maximum values. Weigle and Banks [9] introduced method which treats 3D time-varying dataset as the static 4D data. T-BON technique [11] extends BONO tree [3] for the time-varying datasets. A common BONO tree structure is saved only once for the entire dataset, while minimum and maximum values of the cells are stored separately for each timestep. Time-space Partitioning Tree (TSP) [10] is a standard full octree. Each node of TSP tree has a binary time tree associated. The partial rendered sub-volumes are cached are re-used for faster visualization. The approach of Gregorski [15] builds a hierarchy of diamonds from the original mesh cells. The mesh refinement process (sequence of split and merge operations) ruled by the min, max and error values of the active diamonds is initiated for each iso-surface query, starting from either current refinement or from a root diamond of hierarchy. Recently the Difference Intervals method [21] has been introduced, encoding change of a cell's status by either Add or Remove operation (a cell becomes active/inactive) similarly to the encoding of the video frames.

All of the methods described above assume static simulation mesh. They exploit the fact that the number of mesh cell and their geometry do not change during a simulation. In a dynamic simulation mesh both the number of cells and their geometry change under the deformation of the simulation domain boundaries. Therefore none of the methods above is suitable for the dynamic mesh datasets.

Only a few methods able to handle dynamic simulation mesh were introduced. A system for interactive visualization of the datasets with dynamic simulation mesh is introduced in [18], which assumes certain time intervals in a dataset (topology zones). Within a topology zone the number of cells and their correspondence between timesteps remains constant. Mesh cells are not matched or tracked over topology zone borders (rezone points). The method [22] tries to re-establish inter-timestep cell correspondence based on the geometric and positional similarity of the 2D cells. The method is based on the previous research on the shape reconstruction from planar cross-sections [4]. Later [23] introduced a method that pre-processes all cells into a list of diamonds. Each created diamond is composed of two cells sharing common face. Diamonds are stored in a data structure that allows fast identification of the diamonds intersected by the isosurface for user-defined isovalue.

The method presented in this paper pre-processes the data from each simulation timestep sequentially. First the active cells (cells intersected by the isosurface) are identified. For each active cell we try to find similar cell in the previous timestep and efficiently re-use its geometry information. Similarity of the compared cells is evaluated using our newly proposed metric (Sect. 3). Processed cells are stored in the tree-based structure, which accommodates geometry and scalar values of the cells for all timesteps (Sect. 4).

III. METRICS

The core of our proposed method is a metrics δ for evaluation of temporal changes in a cell's shape. Let's consider two N-dimensional cells C_1 and C_2. The metrics δ compares the shapes of C_1 and C_2 by evaluating of how much the vertices of C_1 have to move in order to *morph* C_1 to C_2:

$$\delta: C_1 \times C_2 \rightarrow \vec{D} \quad (1)$$

where: $\vec{D} \in R^N$, and $|D| \in \langle 0, \infty \rangle$.

For the rest of this article, let's define L to be the length of the longest edge of C_1. The resulting distances between the shapes of C_1 and C_2 have the following meanings:

$|D| = 0$, shape and position of C_1 is equal to C_2
$|D| < 1$, vertices of C_1 moved in total by less than $N * L$
$|D| = 1$, the vertices of C_1 moved by $N * L$ between C_1 and C_2
$|D| > 1$, vertices of C_1 moved in total by more than $N*L$

In other words, if $|D| = 0$ then positions of vertices of C_1 are exactly the same as the positions of vertices of C_2. As the value of $|D|$ grows the two compared cells become more and more different. Note that the metrics δ is not translation, rotation and scale invariant.

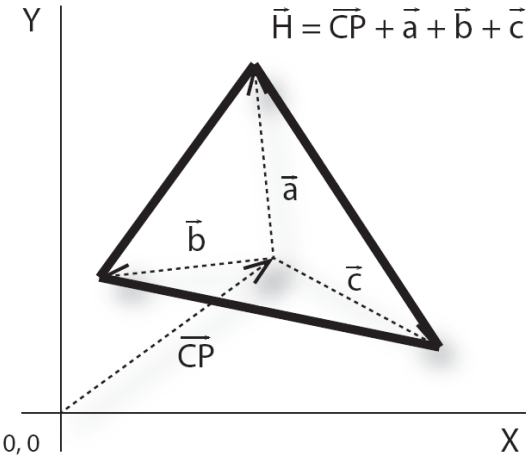


Fig.2: Computation of vector $\vec{H}$ for a 2D triangular cell C. $\vec{H}$ is equal to the vector sum of average center CP of cell C and all vectors from the center CP to the cell's vertices. Computation of $\vec{H}$ scales well to the other dimensions, e.g. tetrahedral cells in 3D. Calculated vector $\vec{H}$ is kept as a description of a cell's position and shape and is used later in the algorithm presented.

The metrics (or distance function) δ between the shape and position of C_1 and C_2 is defined as:

$$\delta(C_1, C_2) = \vec{H}_1 - \vec{H}_2 \quad (2)$$

$$\vec{H}_i = center(\overrightarrow{CP_i}) + \sum_{j=1}^{N} (\vec{V}_{ij} - \vec{C}P_i), i \in \{1,2\} \quad (3)$$

where: $\vec{H}_i$ represents the center point of the i-th cell computed as the average sum of the positions of all cell's vertices; N is dimensionality of the cell and $\vec{V}_{ij}$ represents position of j-th vertex of the i-th cell.

IV. ALGORITHM

Input parameters of our algorithm are start and finish timestep T_S, finish timestep T_F and selected isovalue q. The algorithm presented is thus able to identify active cells for one selected isovalue for any timestep between T_S and T_F of the dataset.

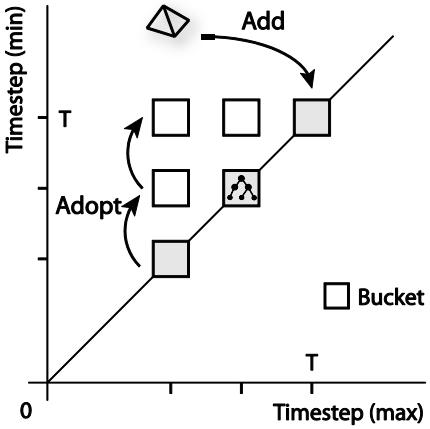


Fig.3: Buckets of the proposed data structure are organized in the 2D grid. Each bucket contains cells organized in a Red-Black tree. The operation Add for timestep T adds a cell into the bucket on position (T, T) of the grid of buckets. The operation Adopt removes the cell from its previous bucket and adds it to the bucket over the original one. In this way, similar cells are re-used between timesteps. For each cell of the dataset either *Add* or *Adopt* is performed.

Next we describe our data structure that keeps all active cells for entire time window between user-selected start and finish timestep. The data structure consists of 2-dimensional lattice of $M \times M$ buckets, where M is equal to the total number of simulation steps. Only half of the buckets above the diagonal in lattice is used. The other half of the buckets under the diagonal plays no role in our algorithm and does not need to be initialized. Along each axis, one bucket represents one timestep. X and Y axis of the lattice have meaning of minimum and maximum timestep similarly to the min-max isovalues in Span-Space [5].

Each bucket in our data structure accommodates certain amount of cells organized in a Red-Black tree. The Red-Black trees introduced by Guibas and Sedgewick in 1978 [1] are self-balancing binary trees, which guarantee worst-case running time $O(n \log n)$, for n accommodated items, for insert, delete and search operations. The data in Red-Black tree are kept in the non-leaf nodes (one data item per node), thus their space complexity is $O(n)$. Space complexity $O(n)$ makes the Red-Black trees suitable data structure for large number of cells that we need to accommodate due to fact that the geometry of simulation mesh is changing with each discrete timestep of a simulation.

Before we describe the algorithm itself, let us define two operations for a cell C: *Add* and *Adopt*, see Fig.3. Operation *Add*] for the cell C and the timestep T, adds C into the bucket (T, T), which lies on the diagonal of the lattice. On the other hand, operation *Adopt* for cell C and timestep T removes C from the bucket $(x, T-1)$ and adds C into the bucket (x, T), where $x \in \langle 0, T-1 \rangle$. By adding or removing a cell from a bucket we mean inserting/deleting the cell out of/into the tree inside the bucket.

The algorithm to populate our data structure is governed by the outer loop in which all timesteps from the user-defined time span T_S, T_F are prepossessed one-by-one, starting from earliest one. For the currently processed timestep, only active cells (e.g. cells intersected by the isosurface) are filtered out and further processed.

For each active cell we keep geometry (position of its vertices), values associated with its vertices and hash key H. The hash key H for a cell is computed using equation Eg.3 from the previous section [24]. H is used later in our algorithm for accommodation of a cell into the buckets of the data structure described above as well as for the comparison of the cell's shape with a candidate cell from successive timestep.

The algorithm starts at the row equal to the first timestep T of the lattice of buckets, see Fig.4. For each timestep all its active cells are processed one-by-one. For each active cell C the algorithm runs the inner loop that traverses buckets of the row $T-1$ in the columns 0 to $T-1$. If a candidate cell C_X is found the algorithm performs the *Adopt* operation for the cell C and the timestep T, otherwise the *Add* operation for the timestep T is performed.

The metrics δ described in the previous section is used for assessment of temporal variations of a cell's shape. C_X (from the timestep $T-1$) is pronounced to be a *predecessor* of C (from the timestep T) only if the distance $|D|$, Eq.1., between C and C_X is smaller than the user defined threshold Delta. In such case the operation *Adopt* is performed, removing C_X from its bucket B and inserting C as an approximate substitution of C_X into the bucket above B. If a candidate C_X for C is not

found, the operation *Add* for the cell C is performed. By increasing the treshold Delta for $|D|$ the user may store active cells more space-efficiently (i.e. more *Adopt* operations will be performed); however, the higher the Delta the higher approximation of the actual isosurface is extracted and rendered. In this way the user can balance between the algorithm's space demands and isosurface accuracy.

During the isosurface rendering step for a timestep T, the active cells are collected from all buckets in and above the grid row corresponding to T and left of and of the column T, see Fig.4. Once the active cells are collected, the isosurface

geometry can be computed and rendered out.

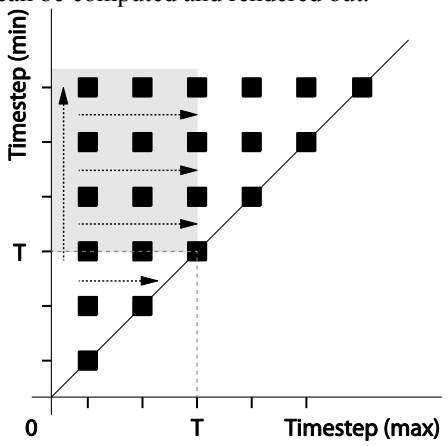


Fig.4: Active cells extraction.

The active cells for the timestep T are collected from the buckets that lie inside or on the borders of the greyed-out area. Dotted arrows show the way in which the buckets are traversed during active cells extraction. For each traversed bucket all its cells are collected by traversing its internal tree.

V. RESULTS

In order to test performance of our method, we ran the series of tests on two datasets produced during Computational Fluid Dynamic simulations. Both datasets are from simulations of the combustion process inside a valve of a diesel engine. Multiple scalar and vector variables like pressure or temperature were computed. The specificity of both datasets is that the geometry of the mesh is not fixed. The total number and shape of the cells changes with each discrete simulation step. The goal of the mesh update procedure is to re-build the mesh around moving parts inside valve. The moving parts are piston and fuel inlets on the top of cylinder.

The Move3D dataset consists of 149 timesteps, capturing a complete combustion cycle inside cylinder. The total number of cells (as well as their shape) varies between 40k and 115k. The dataset occupies 8.31 GB of space. The second dataset Valve consists of 960 timesteps and occupies 21.2 GB of storage. Total number of cells for the second dataset changes between 60k and 90k.

All tests were done on Intel Core2 Duo T5750 2GHz workstation with 4 GB of RAM. The method has been implemented in C language using Standard Template Library and compiled with GCC compiler.

A preprocessing time has been measured for both datasets. Fig.4. shows that the preprocessing time depends on the overall size of the dataset (total number of cells for given isovalue) rather than on the ratio of *Add* and *Adopt* operations performed. This is due to the fact that the time required for *Adopt* operation (and searching for predecessor cell) is only slightly different than the time required for *Add*.

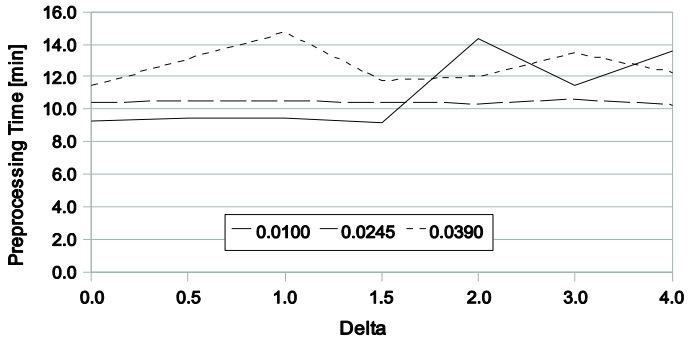


Fig.5: Preprocessing times for Move3D dataset. Relationship between preprocessing time and the Delta parameter of the proposed method. The preprocessing time depends on the total number of cells for given isovalue rather than on the number of Add and Adopt operations performed.

The saving of runtime memory depends on the value of the parameter Delta, Eq. 2. The higher the value of Delta the higher space savings are achieved and vice versa. Higher threshold for Delta means that higher number of cells will be re-used between the timesteps, i.e. operation *Adopt* will be performed instead of *Add*, which requires new memory allocation. Thus, the overall space saving depends on the similarity of the changing simulation mesh in the adjacent timesteps.

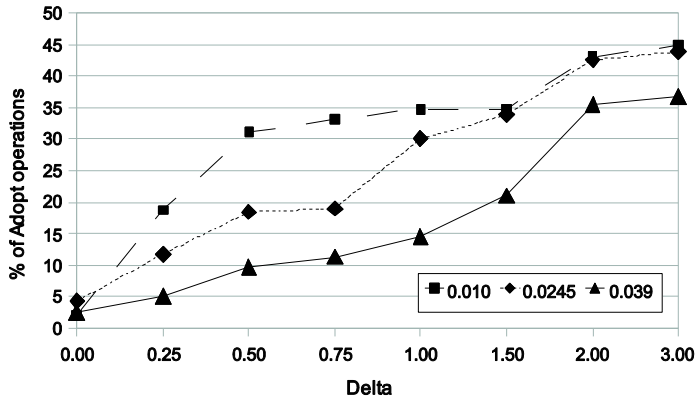


Fig.6: Move3D dataset. Percentage of the Adopt operations vs. parameter Delta of the proposed method for three different isovalues. For high values of Delta the method achieves around 50% of the Adopt operations, which is proportional to the space savings for the given value of Delta.

Fig.5 shows the relationship between the increasing value of the parameter Delta and the percentage of the *Adopt* operations performed over the cells in the Move3D dataset for three different isovalues. We have only chosen parameter Delta in the range 0.0 to 3.0; higher values of Delta show higher degree of the damage of the produced isosurface, Fig.8.

ISBN: 978-1-61804-236-1

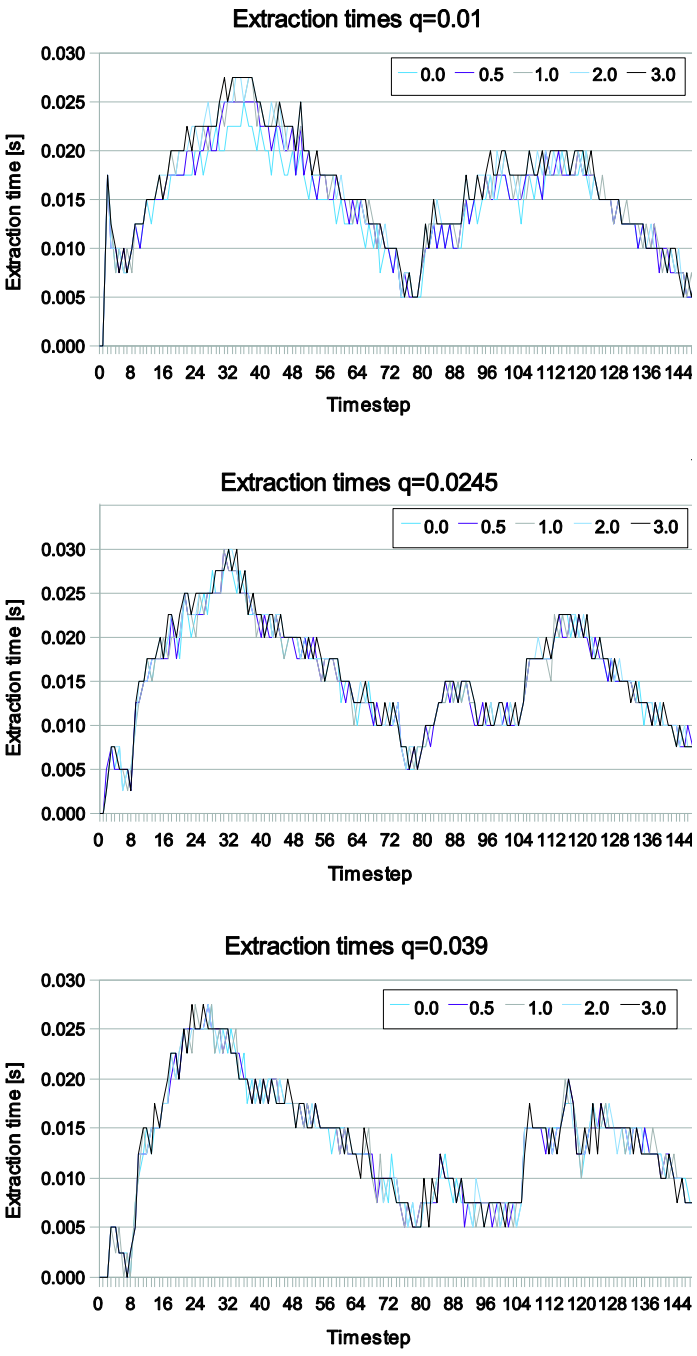


Fig.7: Extraction times for the Move3D dataset for three different isovalues q: 0.01, 0.0245 and 0.039 and different values of the parameter Delta between 0.0 and 3.0.

Finally, we measured the extraction times of the sets of active cells for three different isovalues and different values of parameter Delta of the Move3D dataset. Due to the common difficulties in measuring code execution time under 1 ms, for each pair isovalue-Delta a total of 400 extractions were made and the average time was computed. The overall extraction time of the active cells show no dependency on the parameter Delta used, but is proportional to the number of buckets visited during the extraction step (i.e. total number of active cells extracted). Fig.7 shows extraction times for the Move3D datasets for three different isovalues and parameter Delta between 0.0 and 3.0.

For visual assessment of the damage fragments of the produced isosurface with growing parameter Delta, Fig.8 shows isosufaces for the isovalue 0.0245 of the quantity AMU (total viscosity) of the Move3D dataset for three different values of Delta: 0.0, 1.0 and 3.0. For the tested datasets Delta > 3.0 produced considerable damage of the isosurface. However, the degree of the damages in higher Deltas is dataset-depended and thus left to be a choice of the user.

VI. Conclusion

An efficient method has been described, capable of interactive visualization of the evolving isosurfaces from the time-varying datasets with dynamic mesh. The method consists of the computationally inexpensive preprocessing step with logarithmic space and time complexity and the extraction step, during which the active cells are idenfied and collected. Two basic stones of the method are the metrics for assessment of the temporal change of a cell's shape (Sec. 3) and the buckets-based data structure (Sec.4) that facilitates space-efficient storage of the similar cells.

The method is particularly suitable for the applications where fast insight into the dataset is more important than high accuracy of the produced isosurfaces. The relative simplicity of the proposed method allows its easy implementation.

ACKNOWLEDGMENT

The authors thank to colleagues at the University of West Bohemia for fruitful discussions and to anonymous reviewers for their comments which helped to improve this manuscript.

References

[1] Guibas L. J., Sedgewick R.: *A dichromatic framework for balanced trees*. Proceedings of the 19th Annual Symposium on Foundations of Computer Science, 8-21.

[2] Lorensen W. E., Cline H. E.: Marching cubes: *A high resolution 3D surface construction algorithm*. Proceedings of ACM SIGGRAPH '87, 163-169.

[3] Wilhelms J., van Gelder A.: *Octrees for faster isosurface generation*. ACM Trans. Graph., 11(3), 201-227.

[4] Bajaj Ch. L., Coyle E. J., Lin K.-N.: *Arbitrary topology shape reconstruction from planar cross sections*. Graphical Models and Image Processing, 58(6), 524-543.

[5] Livnat Y., Shen H.-W., Johnson Ch. R.: *A Near Optimal Isosurface Extraction Algorithm Using the Span Space*. IEEE Transactions on Visualization and Computer Graphics, 2(1), 73-84.

[6] Shen H.-W., Hansen Ch. D., Livnat Y., Johnson Ch. R.: *Isosurfacing in Span Space with Utmost Efficiency*, IEEE Visualization '96, 287-294.

[7] Amsden A.A.: KIVA-3V: *A block-structured KIVA program for engines with vertical or canted valves*. Los Alamos NATIONAL LABORATORY, Technical Report LA-13313-MS.

[8] Shen H.-W.: *Iso-surface extraction in time-varying fields using a temporal hierarchical index tree*. Proceedings of Visualization '98, 159-166.

[9] Weigle Ch., Banks D. C.: *Extracting iso-valued features in 4-dimensional scalar fields*. Proceedings of IEEE Symposium on Volume Visualization '98, 103-110.

[10] Shen H.-W., Chiang L.-J., Ma K.-L.: *A fast volume rendering algorithm for time-varying fields using a time-space partitioning TSP tree.* Proceedings of Visualization '99, 371-377.

[11] Sutton P., Hansen Ch. D.: *Isosurface extraction in time-varying fields using a temporal branch-on-need tree*. Proceedings of Visualization '99, 147-153.

[12] de Leeuw W., van Liere R.: *Chromatin decondensation: a case study of tracking features in confocal data*. Proceedings of Visualization '01, 441-444.

[13] Fluent news: *Dynamic Mesh*, Volume: XI, editor: Liz Marshall, Fluent Inc.

[14] Donea J., Huerta A., Ponthot J.-Ph., Rodriguez-Ferran A.: *Encyclopedia of Computational Mechanics*, Volume 1. John Wiley \& Sons.

[15] Gregorski B.: *Adaptive Extraction of Time-Varying Isosurfaces*. IEEE Transactions on Visualization and Computer Graphics 10(6), 683-694.

[16] Cavallo P., Hosangadi A., Ahuja V.: *Transient simulations of valve motion in cryogenic systems*. Proceeding of 35th AIAA Fluid Dynamics Conference and Exhibit.

[17] Szymczak A.: *Subdomain-aware contour trees and contour tree evolution in time-dependent scalar fields*. Proceedings of Shape Modeling International '05, 136-144.

[18] Doleisch H., Mayer M., Gasser M., Priesching P., Hauser H.: *Interactive Feature Specification for Simulation Data on Time-Varying Grids*. SimVis'05, 291-304.

[19] Waters K. W., Co Ch. S., Joy K. I.: *Isosurface Extraction Using Fixed-Sized Buckets*. IEEE VGTC Symposium on Visualization, 207-214.

[20] Bernardon F., Callahan S., Comba J., Silva C.: *Interactive volume rendering of unstructured grids with time-varying scalar fields*. Proceedings of Eurographics Symposium on Parallel Graphics and Visualization '06, 51-58.

[21] Waters K. W., Co Ch. S.: *Using Difference Intervals for Time-Varying Isosurface Visualization*. IEEE Transactions on Visualization and Computer Graphics, 12(5), 1275-1282.

[22] Petrik S., Skala V.: *Iso-contouring in Time-varying Meshes*. SCCG 2007 Proceedings, 216-223.

[23] Petrik S., Skala V.: *Z-Diamonds: A Fast Iso-surface Extraction Algorithm for Dynamic Meshes*. IADIS Computer Graphics and Visualization proceedings 2007.

[24] Hradek,J., Skala,V.: *Hash Function and Triangular Mesh Reconstruction*, Vol.29, No.6., pp.741-751, Computers&Geosciences, Pergamon Press, ISSN 0098-3004, 2003

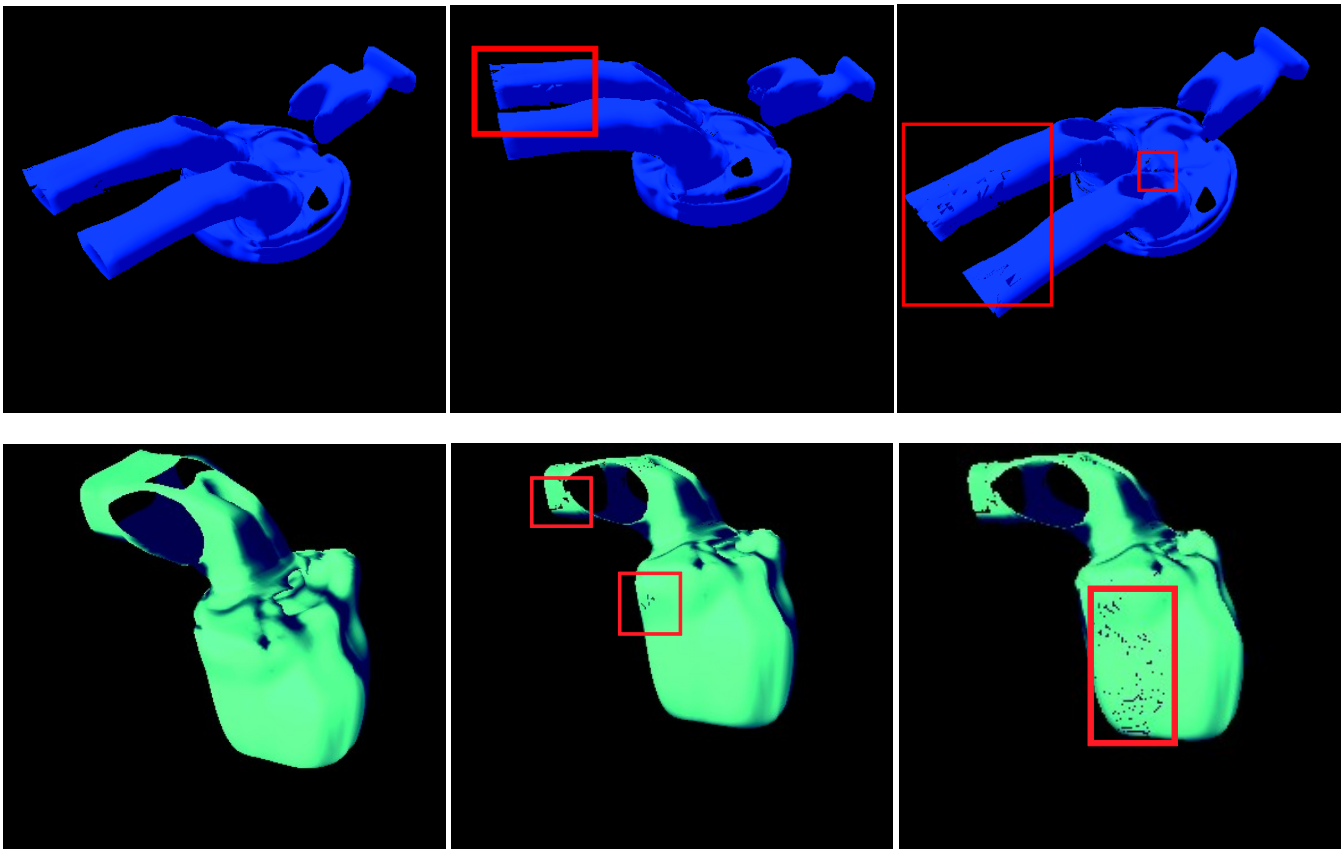


Fig.8: Isosurfaces of the Move3D dataset. Top row: timestep 13, bottom row: timestep 113. Isovalue q=0.0245 of the AMU (total viscosity).
Each isosurface has been generated with different value of Delta; from left to right: Delta = 0.0, 1.0, 3.0.

Computer Vision Applied for Accessing to Machine Information using Sobel Operator

Chávez S. Rodolfo, Lozano C. Ruben, and Pedraza M. Luis

***Abstract*—**This paper describes an application of artificial intelligence, consists of hardware and software that is responsible for making decisions by comparing images taken previously and are stored in the program developed, it is important calibrate the device with the user´s eyes. The operator used to process the image is the Sobel algorithm, this algorithm is used in image processing, especially in the edge-detection algorithms. In theory it is a discrete differential operator that computes an approximation to the slope of the function of the intensity of an image. This application is an approach to the meaning of a neural network, stored as arrays of measurement standards and the choice of one of them by comparison with the sample in real time.

***Keywords*—** Images processing; carrier sobel; images filters; gradient; operator differential.

I. Introduction

THE first step in any image analysis is segmentation. By segmentation divides the image into parts or objects that comprise it. The level at which the subdivision is done depends on the particular application, when finished segmentation detected all objects of interest to the application. In general, the automatic segmentation is one of the most complicated tasks in the processing image. The segmentation will ultimately lead to success or failure the process of analysis. In most cases, a good segmentation give rise to a right solution, so that should make every effort possible segmentation stage.

The image segmentation algorithms are generally based on two basic properties of gray levels image: discontinuity and similarity. Within the first category is trying to divide the image based the abrupt changes in gray level. The areas of interest in this category are the detection of points, lines and edges in the image. The areas within the second category are based on the techniques of threshold, growth regions, and division and fusion techniques

Among the numerous applications for edge detection, digital artists use it to create stunning images with contours as the output of an edge detector can be added to an original image to highlight the edges. Edge detection is often the first step in image segmentation, which is a field of image analysis, and is used to group pixels into regions to determine a composition of the image. Edge detection is also used in image registration, which aligns two images could be acquired at separate and different sensors.

The edges have the objects, their shape, size and also on its texture. The edges are in areas of an image where the intensity level fluctuates sharply, the faster the change of intensity, the axis or edge is stronger.

In general, the edges of objects in an image can distinguish the more or less abrupt changes in value between two or more adjacent pixels. Can make a general classification of the edges as address:

- Vertical-Edges, when connected pixels vertically, and have different values from previous or later.
- Horizontal-Edges, when connected pixels horizontally, and these have different values from previous or later.
- Tapering, when have a combination of horizontal and vertical components.

The difference between the values of the pixels indicates the sharp edge, so that major differences have marked edges and lower edges have softened.

The filters used for edge detection are differential filters, which are based on the derivation or differentiation. Since the average of the pixels in a region tends to blur or soften the details and edges of the image. [1]

The main objective of this project is to design and implement a prototype to perform a scan and parameterization of the retina to handle any program based on the Windows platform by moving the mouse cursor, for which they have raised the following specific objectives:

Construction of a hardware that generates a controlled environment for the acquisition of images, implement an interface to the PC for transferring images, implement image processing algorithms to highlight and draw contours, implement a Visual Basic 6.0 software responsible for processing the images and that depending on the position of the eye move the mouse cursor.

You want to leave a knowledge base for future research in the field of Artificial Vision.

II. Development and Construction

The project is developed in 5 phases:

Rodolfo Chavez is with the Research Group GIDENUTAS, "Universidad Distrital Francisco José de Caldas" Bogotá DC., Colombia. (phone: 57-321-3109789; e-mail: rechavezs@correo.udistrital.edu.co).

Hardware:

A. Construction of the controlled environment.
B. Implementation of the interface with PC.

Software:

C. Image Acquisition.
D. Digital image processing.

A. *Controlled environment*

The importance of the controlled environment is to allow the acquisition of an excellent image of the human eye under certain lighting conditions and texture that allows the application of processes such as edge detection and contour and thus can ensure the efficient operation of the software.

The built a lens, which was taken from two industrial safety masks (Figure 1), which were adapted for positioning the left eye, the atmosphere is totally closed to prevent any unwanted light leakage, segments opaque image because of darkness, shadows or too bright.

Fig.1 Mask Industrial Safety

B. *With Pc Interface Implementation*

The lens is 5.50 cm in diameter, black to isolate any penetration of light and prevent see the module at the bottom, side and completely covered black matt, looking for uniformity and light absorption, which originates through 3 led's white light jet (3 mm) in two different directions appropriate to diffuse the light, provided that at no time point directly to the eye or the camera lens so as not to cause noise in the digital processing image.

Is used a Web camera (located at the bottom of the lens) with a resolution of 640 x 480 pixels (Figure 2), video format of 32 bits, with manual focus lens type with USB communication protocol.

Fig.2 a) Lens appropriate, b) Final prototype.

C. *Image Acquisition*

The image is captured by the webcam with a resolution of 640 X 480 pixels in BMP format, this image is not stored directly to the PC but the software works with real-time image acquisition of the image is captured. First of all take the initial image, (Top, Left, Right, Down, Center) previously stored in the software in order to work in the processing of each image taken earlier by the webcam.
(Figure 3)

Positions:

1. *Up*: The eye is in the up position, the cursor will move up.
2. *Down*: The eye is in the down position, the cursor will moves down.
3. *Right*: The eye is in the right position, the cursor will move to the right.
4. *Left*: The eye is in the up position, the cursor will move up.
5. *Center*: The eye is in the center position, the cursor will not move.

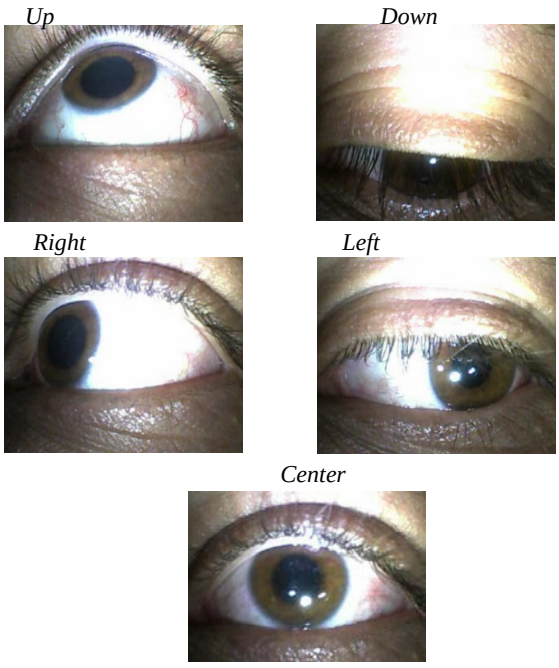


Fig. 3 Five positions of the Human Eye

ISBN: 978-1-61804-236-1

D. Digital Processing

A color model is a method for specifying colors in some standard way. It generally consists of a 3D coordinate system and a subspace of that system in which each color is represented by a single point. In RGB, each color is represented as 3 values R, G and B, indicating the amounts of red, green and blue which make up the color. In HSV, the "true color" attributes (red, green, blue, orange, yellow, and so on). The amount by which the color has been diluted with white. The whiter in the color, the lower the saturation. The degree of brightness a well it color has high intensity; a dark color has low intensity.

For this process studied various methods, to choose the best before applying any differential operator and obtain the desired final image, are the following:

1. Grayscale

In grayscale images, however, we do not differentiate how much we emit of the different colors, we emit the same amount in each channel. What we can differentiate is the total amount of emitted light for each pixel; little light gives dark pixels and much light is perceived as bright pixels.

When converting an RGB image to grayscale, we have to take the RGB values for each pixel and make as output a single value reflecting the brightness of that pixel. One such approach is to take the average of the contribution from each channel: (R+B+C)/3. However, since the perceived brightness is often dominated by the green component, a different, more "human-oriented", method is to take a weighted average:
0.3R + 0.59G + 0.11B.

When making or taking a photo in grayscale may represent a set of colors in tones of gray, or even put each color intensity. The gray scales are different from black and white photographs in which the colors are coded in white or black, gray scale provides a range of shades of gray in between.[2]

The grayscale images, use 8 bits to represent each pixel which allows only one scale with 256 intensities (or gray scale), 2 possible values for each bit (0 and 1) raised to 8 bit used to represent each pixel, gives us 256 different color shades that can be represented in a grayscale image. [3]

In this case 0 to 256 in grayscale using a threshold greater than or equal to 70 for the detection of edges, mean a 1 (one) in the image.

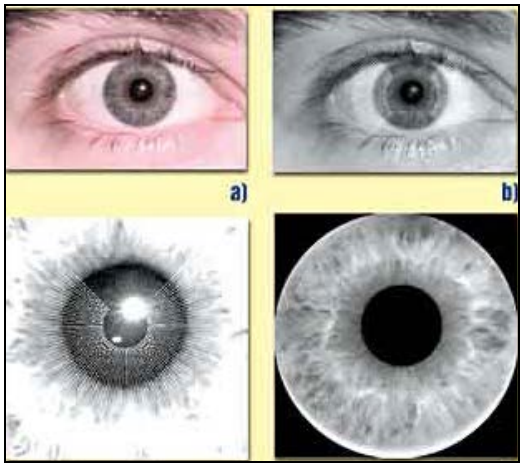


Fig. 4 Grayscale Image. a) Original image. b) Grayscale.

See also figure 5, grayscale colors reversed. This option is used but depending the needs.

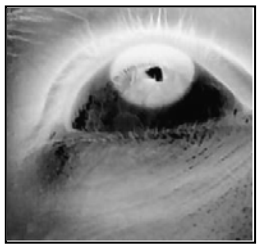


Fig. 5 Colors reversed in grayscale.

For this preliminary stage was chosen without investing grayscale colors (Figure 3).

2. Edge Detection

The idea behind most of the edge detection techniques is the calculation of a local operator bypass since a pixel belongs to an edge if there is an abrupt change between levels of gray with its neighbors. [4]

One of the most important and simple process is to *detect edges*. Important because it can start extracting important image information, such as the shapes of the objects that compose it, and simple because the edge detection operators are simple convolution masks. These operators are used in applications for pattern recognition, industrial, military, etc..

Among the numerous applications for edge detection, digital artists use it to create stunning images with contours as the output of an edge detector can be added to an original image to highlight the edges. Edge detection is often the first step in image segmentation, which is a field of image analysis, and is used to group pixels into regions to determine a composition of the image. Edge detection is also used in image registration, which aligns two images could be acquired at separate and different sensors.

ISBN: 978-1-61804-236-1

The *edges of image* contain much information thereof. The edges have where the objects, their shape, size, and also on its texture. The axes or edges are in areas of an image where the intensity level fluctuates sharply, the faster the change of intensity, the axis or edge is stronger. A good edge detection process facilitates the development of the boundaries of objects with which the object recognition process is simplified. To detect the edges of objects, detected those points that form the edge.[6] - [9].

3. Link Edges

The discontinuity detection techniques ideally provide the pixels corresponding to the contours or boundaries between regions image. In practice this set of pixels is usually not fully characterize these contours due to the presence of noise, disruption in their own contours due to uneven lighting and other effects that introduce spurious discontinuities in intensity. In general, after the 8 procedures edge detection techniques are often used linked or other contour detection techniques designed to unite the pixels significant contour edges.

Common the gradient (or gradient orthogonal) are horizontal and vertical edges. These operators work by convolution. The operators of *Prewitt, Sobel, Roberts and Frei-Chen* are double or two stages. Edge detection is performed in two steps, first applying a mask to find horizontal edges, and the second step is seek the vertical, the end result is the sum of both.[5] Are some common convolution masks below. Detectors row (horizontal) are *Hh* detectors and column (vertical) are *Hv*:

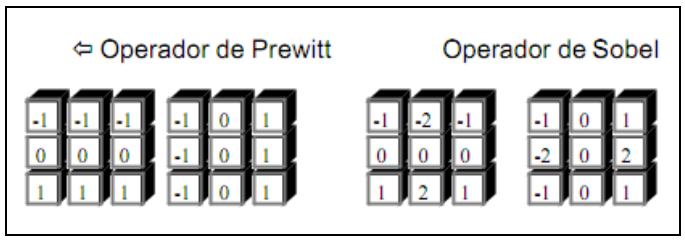


Fig. 6 Masks of Prewitt and Sobel to detect edges

Figure 6 shows how each of the masks highlights a type of edge, depending on their orientation. Differential filters help to detect edges as areas in the original image are of a uniform tone (whatever) middle gray (values close to zero) becomes. Meanwhile, the edges, areas where there is an abrupt change in intensity, are emphasized. Some are black (negative values and other targets (positive values).

Ideally, edge detection techniques yield pixels lying only on the boundaries between regions. In practice, this pixel set seldom characterizes a boundary completely because of noise breaks in the boundary due to non-uniform illumination.
Thus, edge detection algorithms are usually followed by linking and other boundary detection procedures designed to assemble edge pixels.

Other edges are emphasized and are gray (values close to zero) [7][10].

This alteration of the edges produces an illusion of relief. The image appears to sink and stand, lit by a source of light. Rinses appear to be more enlightened, and the shadows seem obscured. The areas in the original image were darker seem to sink, while lighter areas appear to excel.

Fig. 7 Binary image with Sobel operator

a) Math Formulation and equations

It is a discrete differentiation Operator which computes an approximation of the gradient of the image intensity function. The Sobel operator is based on convolving the image with a small, separable, and integer valued filter in horizontal and vertical direction and is therefore relatively inexpensive in terms of computations. The sobel operator calculates the gradient of the image intensity at each point, giving the direction of the largest possible increase from light to dark and the rate of change in that direction. Each image point, the gradient vector points in the direction of largest possible intensity increase, and the length of the gradient vector corresponds to the rate of change in that direction.

Mathematically, the operator uses two 3 × 3 kernels of elements to apply convolution to the original image to calculate approximations of derivatives, a kernel for horizontal changes and one for vertical. If defined as the original image are the two kernels for each point representing horizontal and vertical approaches of the derivatives of intensity [7][8], the result is calculated as:

$$\mathbf{G_x} = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} * \mathbf{A} \quad \text{and} \quad \mathbf{G_y} = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * \mathbf{A} \qquad (1)$$

Horizontal *Vertical*

At each point of the image, the results of approximations of the horizontal and vertical gradients can be combined to obtain the magnitude of the gradient [10] by:

$$\mathbf{G} = \sqrt{\mathbf{G_x}^2 + \mathbf{G_y}^2} \qquad (2)$$

ISBN: 978-1-61804-236-1

Where,

G = SOBEL
Gx = Horizontal Edges
Gy = Vertical Edges

This operator can be seen in the example of Figure 8.

Typically, an approximate magnitude is computed using:

$$|G| = |Gx| + |Gy| \quad (3)$$

The angle of orientation of the edge (relative to the pixel grid) giving rise to the spatial gradient is given by:

$$\theta = \arctan(Gy/Gx) \quad (4)$$

In this case, orientation 0 is taken to mean that the direction of maximum contrast from black to white runs from left to right on the image, and other angles are measured anti-clockwise from this.

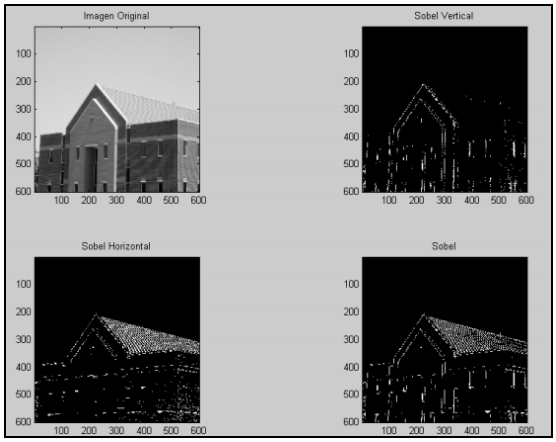


Fig. 8 Example of sobel operator applied to grayscale image.

Please submit your manuscript electronically for review as e-mail attachments.

III. Design and Implementation of Control Software

To design the control software implemented three phases or windows:

A. Calibration Program.
B. Sampling Program.
C. Mouse Function

A. Calibration Program

In this window you will find two viewers to be launched, a playback device and the other picture taken for calibration, also have twenty-five buttons to make the respective pictures; 5 for each position, as shown in Figure 9:

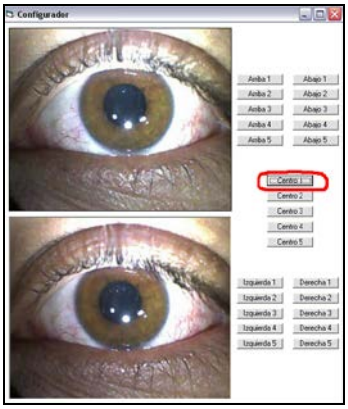


Fig. 9 Calibration Window

B. Sampling Program

After having all the pictures calibration stored in a folder that is responsible for organizing software can close this window and proceed to the sampling program, there are two viewers, the first device will playback in real time and in the second display will be shooting burst by the device, the image taken with 100fps is stored as an image input to the digital image processing, the window as shown in Figure 10:

Fig. 10 Sampling Window

C. Mouse Function

Finally the third window will notice the digital processing of the image and thus recognition of eye position, indicating a viewer will appear in words the position and move the mouse in the position previously indicated by the eye movement the mouse was in short sections for increased accuracy in the displacement of the user on the computer (Figure 11).

For digital image processing from the device was done using the ezVidCap control. This control allows you to capture images from a video device to the computer. The communication connection from the camera to the PC's USB

control ezVidCap looking at the root of the controller drivers from the computer device and displays video image capture in real time on a picture (picture box) within VISUAL BASIC 6.0.

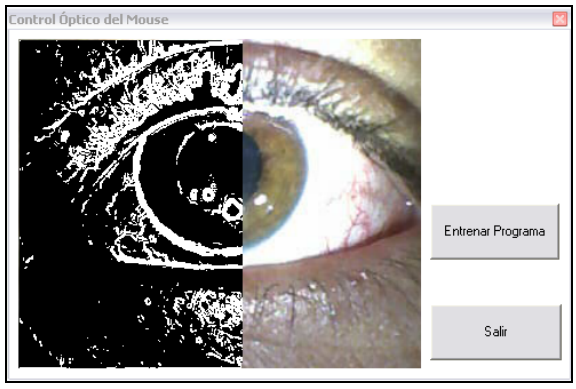


Fig. 11 Final comparation and position of window

Finally, through a process of comparison matrices (result of operator) is known the position of the eye, which is implemented through a function SetCursorPos, routine for pointer movement.

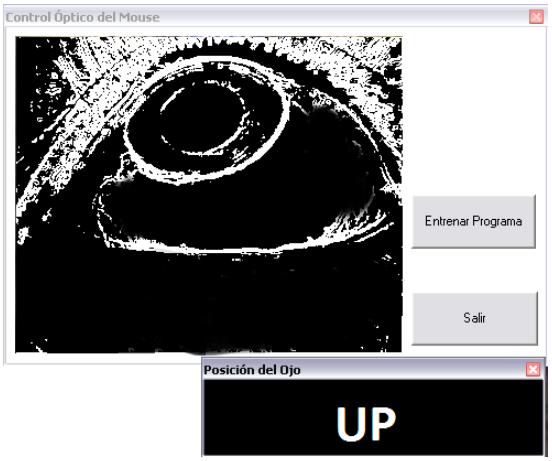


Fig. 12 Running Program

The immediate application of the system in principle would be to improve communication for people with severe motor problems (quadriplegia), with very encouraging results. But it also could be implemented in: control of industrial processes (replacement of use of the hands of operators in critical tasks), virtual reality (combining the system with virtual reality technology, it could increase the level of realism of the virtual environment and many applications that arise from public knowledge of this technology.

IV. DISCUSSION

The Sobel edge detector can also be applied to range images like. Edge detection using gradient operators approach tends to work well in cases that involve transitions of intensity images with clearly defined and relatively low noise. Sobel operator gets a good result compared to other operators such as the Laplacian, Prewitt or Frei-Chen, detects edges in all directions and does not increase the noise, but requires a lot of operations and time consuming. The Sobel operator is not as sensitive to noise as the Roberts Cross operator, it still amplifies high frequencies

ACKNOWLEDGMENT

The work is developed through the cooperation of research groups "DIGITI" and "GIDENUTAS" of the Universidad Distrital Francisco José de Caldas of Bogota D.C. (Colombia).

REFERENCES

[1] Department of Electronic Engineering, Telecommunications and Automation. Area Systems and Automation Engineering Academic Year 2005/2006.

[2] The Research and Implementation on Edge Detection Algorithms of Ochotona Curzoniae's Image Based on Grayscale Morphology. *IEEE International Conference on Computational and Information Science.*, 978-0-7695-4501-1/11, 2011.

[3] Converting a Digital Color Photo Into Black and White, includes a background on color filter use in traditional film photography, how black and white conversion works, and a comparison of digital conversion techniques

[4] Advanced methods in segmentation of images in grayscale image processing software lab, 2006.

[5] A Classified and Comparative Study of Edge Detection Algorithms, proceedings of the International Conference on Information Technology: Coding and Computing *IEEE* (ITCC02), Mohsen Sharifi, Mahmoud Fathy, Maryam Tayefeh Mahmoudi, 2002.

[6] R. Haralick and L. Shapiro, "Survey: Image segmentation techniques," Comput. Vis. Graph. Image Processing, vol. 29, no. 1, pp. 100–132, 1985.

[7] Kuzina. T. Y. 2000. Investigación e implementación de lo s algoritmos del campo de visión dinámica por computadora. Tesis Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales. Escuela de Ingeniería. Universidad de las Américas – Puebla, Mayo.

[8] S. Zucker, "Survey region growing: Childhood and adolescence," *Comput.Graph. Image Processing*, vol. 5, no. 4, pp. 382–399, 1976.

[9] A transform for multiscale image segmentation by integrated edge and region detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 1211–1235, Dec. 1996.

[10] Edge Detection of Images Based on Improved Sobel Operator and Genetic Algorithms, IEEE, Zhang Jin-Yu, Chen Yan, Huang Xian-Xiang Xi'an Research Institute of High-tech, Xi'an, Shaanxi, China, 978-1-4244-3986-7/09, 2009.

Developing Flexible Applications with Actors

Agostino Poggi

***Abstract*—**the development of scalable and efficient applications requires some appropriate models and programming technique. This paper presents an actor-based software framework that allows the development of scalable and efficient applications through the possibility of using different implementations of the components that drive the execution of their actors. In particular, the paper shows how the performance of applications can be optimized by choosing the best combination among the alternative implementations of its components.

***Keywords*—** Actor model, software framework, distributed system, Java.

I. Introduction

The availability of large-scale, dynamic, and heterogeneous networks of computational resources and the advent of multi-cores computers make possible the development of high performance and scalable computationally intensive applications. However, the building of such a kind of applications requires the availability of appropriate distributed and concurrent programming models and techniques.

Message passing is the most attractive solution because it is a concurrent model that is not based on the sharing of data and so its techniques can be used in distributed computation too. One of the well-known theoretical and practical models of message passing is the actor model [1]. Using such a model, programs become collections of independent active objects (actors) that exchange messages and have no mutable shared state. Actors can help developers to avoid issues such as deadlock, live-lock and starvation, which are common problems for shared memory based approaches. There are a multitude of actor oriented libraries and languages, and each of them implements some variants of actor semantics. However, such libraries and languages use either thread-based programming, which makes easy the development of programs, or event-based programming, which is far more practical to develop large and efficient concurrent systems, but also is more difficult to use.

This paper presents an actor based software framework, called CoDE (Concurrent Development Environment), that has the suitable features for both simplifying the development of large and distributed complex systems and guarantying scalable and efficient applications. The next section introduces related work. Section 3 describes the software framework. Section 4 details the features that make CoDE suitable for using actor-based systems in different kinds of applications and presents the experimentation results. Finally, section 5 concludes the paper by discussing its main features and the directions for future work.

This work was supported in part by the Ministry of Education, Universities and Research.

A. Poggi is with the University of Parma, Parma, 43100 Italy (phone: +39-0521-905728; fax: +39-0521-905728; e-mail: Agostino.poggi@unipr.it).

II. Related Work

Several actor-oriented libraries and languages have been proposed in last decades and a large part of them uses Java as implementation language. The rest of the section presents some of the most interesting works.

Salsa [2] is an actor-based language for mobile and Internet computing that provides three significant mechanisms based on the actor model: token-passing continuations, join continuations, and first-class continuations. In Salsa each actor has its own thread, and so scalability is limited. Moreover, message-passing performance suffers from the overhead of reflective method calls.

Kilim [3] is a framework used to create robust and massively concurrent actor systems in Java. It takes advantage of code annotations and of a byte-code post-processor to simplify the writing of the code. However, it provides only a very simplified implementation of the actor model where each actor (called task in Kilim) has a mailbox and a method defining its behavior. Moreover, it does not provide remote messaging capabilities.

Scala [4] is an object-oriented and functional programming language that provides an implementation of the actor model unifying thread based and event based programming models. In fact, in Scala an actor can suspend with a full thread stack (receive) or can suspend with just a continuation closure (react). Therefore, scalability can be obtained by sacrificing program simplicity. Akka [5] is an alternative toolkit and runtime system for developing event-based actors in Scala, but also providing APIs for developing actor-based systems in Java. One of its distinguishing features is the hierarchical organization of actors, so that a parent actor that creates some children actors is responsible for handling their failures.

Jetlang [6] provides a high performance Java threading library that should be used for message based concurrency. The library is designed specifically for high performance in-memory messaging and does not provide remote messaging capabilities.

AmbientTalk [7] is a distributed object-oriented programming language that is implemented on an actor-based and event driven concurrency model, which makes it highly suitable for composing service objects across a mobile

ISBN: 978-1-61804-236-1

network. It provides an actor implementation based on communicating event loops [8]. However, each actor is always associated with its own JVM thread and so it limits the scalability of applications on the number of actors for JVM.

III. Software Framework

CoDE (Concurrent Development Environment), is an actor based software framework that has the goal of both simplifying the development of large and distributed complex systems and guarantying an efficient execution of applications.

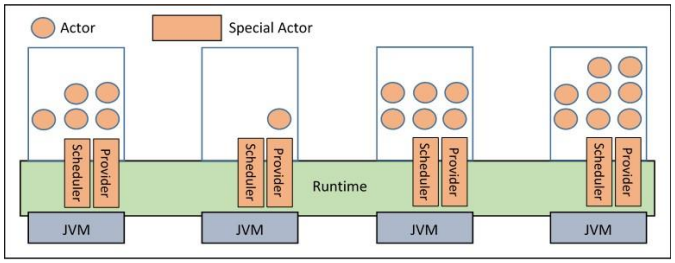


Fig. 1 Application architecture

CoDE is implemented by using the Java language and takes advantage of preexistent Java software libraries and solutions for supporting concurrency and distribution. CoDE has a layered architecture composed of an application and a runtime layer. The application layer provides the software components that an application developer needs to extend or directly use for implementing the specific actors of an application. The runtime layer provides the software components that implement the CoDE middleware infrastructures to support the development of standalone and distributed applications.

A. Model View

In CoDE an application is based on a set of interacting actors that perform tasks concurrently. An actor is an autonomous concurrent object, which interacts with other actors by exchanging asynchronous messages. Moreover, it can create new actors, update its local state, change its behavior and kill itself.

Communication between actors is buffered: incoming messages are stored in a mailbox until the actor is ready to process them. Each actor has a system-wide unique identifier called its address that allows it to be referenced in a location transparent way. An actor can send messages only to the actors of which it knows the address, that is, the actors it created and of which it received the addresses from other actors. After its creation, an actor can change several times its behavior until it kills itself. Each behavior has the main duty of processing a set of specific messages through a set of message handlers called cases. Therefore, if an unexpected message arrives, then the actor mailbox maintains it until a next behavior will be able to process it.

An actor can set a timeout for waiting for the next message and then execute some actions if the timeout fires. However, it has not explicit actions for monitoring the firing of such a timeout: its implementation autonomously observes the firing of the timeout and then executes the actions for its management.

Depending on the complexity of the application and on the availability of computing and communication resources, one or more actor spaces can manage the actors of the application. An actor space acts as "container" for a set of actors and provides them the services necessary for their execution. In particular, an actor space takes advantages of two special actors: the scheduler and the service provider. The scheduler manages the concurrent execution of the actors of the actor space. The service provider enables the actors of an application to perform new kinds of action (e.g., to broadcast a message or to move from an actor space to another one). Fig. 1 shows a graphical representation of the architecture of a CoDE distributed application.

B. Implementation View

An actor can be viewed as a logical thread that implements an event loop [7][8]. This event loop perpetually processes events that represent: the reception of messages, the behavior exchanges and the firing of timeouts. An actor manages the life cycle of the actor by initializing its behaviors, by processing the received messages and the firing of message reception timeouts, and by moving it from a behavior to another one. CoDE provides different actor implementations and the use of one or of another implementation represents one of the factors that mainly influence the attributes of the execution of an application. In particular, actor implementations can be divided in two classes that allow to an actor either to have its own thread (from here named active actors) or to share a single thread with the other actors of the actor space (from here named passive actors). Moreover, the implementation of an actor takes advantages of other five main components: a reference, a mailer, a behavior, a state and an execution manager. Fig. 2 shows a graphical representation of the architecture of an actor.

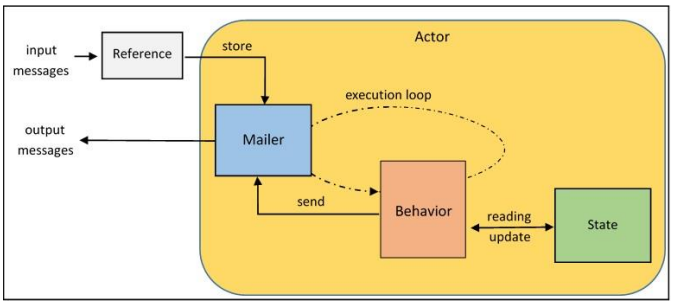


Fig. 2 Actor architecture

A reference supports the sending of messages to the actor it represents. Therefore, an actor needs to have the reference of another actor for sending it a message. In particular, an actor has have the reference of another actor if either it created such an actor (in fact, the creation method returns the reference of the new actor) or it received a message from such an actor (in fact, each message contains the reference of the sender) or whose content enclosed its reference.

A reference has an attribute, called actor address, that

allows to distinguish itself (and then the actor it represents) from the references of the other actors of the application where it is acting. To guarantee it and to simplify the implementation, an actor space acts as "container" for the actors running in the same Java Virtual Machine (JVM) and an actor address is composed of an actor identifier, an actor space identifier and the IP address of the computing node. In particular, the actor identifier is different for all the actors of the same actor space, and the actor space identifier is different for all the actor spaces of the same computing node.

A mailer provides a mailbox for the messages sent to its actor until it processes them, and delivers its messages to the other actors of the application. As introduced above, a behavior can process a set of specific messages leaving in the mailbox the messages that is not able to process. Such messages remain into the mailbox until a new behavior is able to process them and if there is not such a behavior they remain into the queue for all the life of the actor. A mailbox has not an explicit limit on the number of messages that can maintain. However, it is clear that the (permanent) deposit of large numbers of messages in the mailboxes of the actors may reduce the performances of applications and cause in some circumstances their failure.

The original actor model associates a behavior with the task of messages processing. In CoDE, a behavior can perform three kinds of tasks: its initialization, the processing of messages and the management of message reception timeouts. In particular, a behavior does not directly process messages, but it delegates the task to some case objects, that have the goal of processing the messages that match a specific (and unreplaceable) message pattern.

Often the behaviors of an actor need to share some information (e.g., a behavior may work on the results of the previous behaviors). It is possible thank to a state object. Of course, the kind of information that the behaviors of an actor need to share depends on the type of tasks they must perform in an application. Therefore, the state of an actor must be specialized for the task it will perform.

A message is an object that contains a set of fields maintaining the typical header information and the message content. Moreover, each message is different from any other one. In fact, messages of the same sender have a different identifier and messages of different senders have a different sender reference.

A message pattern is an object that can apply a combination of constraint objects on the value of all the fields of a message. CoDE provides a set of predefines constraints, but new ones can be easily added. In particular, one of such constraints allows the application of a pattern to the value of a message field. Therefore, the addition of field patterns (the current implementation offer only a regular expression pattern) will allow the definition of sophisticated filters on the values of all the message fields and in particular on the content of the message.

An actor has not direct access to the local state of the other actors and can share data with them only through the exchange of messages and through the creation of actors. Therefore, to avoid the problems due to the concurrent access to mutable data, both message passing and actor creation should have call-by-value semantics. This may require making a copy of the data even on shared memory platforms, but, as it is done by the large part of the actors libraries implemented in Java, CoDE does not make data copies because such operations would be the source of an important overhead. However, it encourages the programmers to use immutable objects (by implementing as immutable all the predefined message content objects) and delegates the appropriate use of mutable object to them.

An actor space has the duty of supporting the execution of the actions of its actors and of enhancing them with new kinds of action. To do it, an actor space takes advantage of some main runtime components (i.e., factory, dispatcher and registry) and of the two special actors: the scheduler and the service provider.

The factory has the duty of creating the actors of the actor space. In particular, it also creates their initial behavior, chooses their most appropriate execution manager and delegates the creation of their references to the registry.

The dispatcher has the duty of supporting the communication with the other actor spaces of the application. In particular, it creates connections to/from the other actor spaces, maps remote addresses to the appropriate output connections, manages the reception of messages from the input connections, and delivers messages through the output connections. This component works in collaboration with another component called connector.

A connector has the duty of opening and maintaining connections toward all the other actor spaces of the application. In particular, the connector of one of the actor spaces of the application plays the role of communication broker and has the additional duty of maintaining the information necessary to a new actor space for creating connections towards the other actor spaces of the application.

The registry supports the work of both the factory and the dispatcher. In fact, it creates the references of the new actors and supports the delivery of the messages coming from remote actor by proving the reference of the destination actor to the dispatcher. In fact, as introduced in a previous section an actor can send a message to another actor only if it has its reference, but while the reference of a local actor allows the direct delivery of messages, the reference of a remote actor delegates the delivery to the dispatchers of the local and remote actor spaces.

The scheduler is a special actor that manages the execution of the actors of an actor space. Of course, the duties of a scheduler depend on the type of execution manager and, in particular, on the type of threading solutions associated with the actors of the actor space. In fact, while Java runtime environment mainly manage the execution of active actors, CoDE schedulers completely manage the execution of passive

actors.

The service provider is a special actor that offers a set of services for enabling the actors of an application to perform new kinds of actions. Of course, the actors of the application can require the execution of such services by sending a message to the service provider. In particular, the current implementation of the software framework provides services for supporting the broadcast of messages, the exchange of messages through the "publish and subscribe" pattern, the mobility of actors, the interaction with users through emails and the creation of actors (useful for creating actors in other actor spaces).

Moreover, an actor space can enable the execution of an additional runtime component called logger. The logger has the possibility to store (or to send to another application) the relevant information about the execution of the actors of the actor space (e.g., creation and deletion of actors, exchange of messages, processing of messages and timeouts, exchange of behaviors). The logger can provides both textual and binary information that can be useful for understanding the activities of the application and for diagnosing the causes and solving the possible execution problems. Moreover, the binary information contain real copies of the objects of the application (e.g., messages and actor state); therefore, such an information can be used to feed other applications (e.g., monitoring and simulation tools).

Finally, the actor space provides a runtime component, called configurator, which simplifies the configuration of an application by allowing the use of either a declarative or a procedural method (i.e., the writing of either a properties file or a code that calls an API provided by the configurator).

IV. Configuration Guides

The quality of the execution of a CoDE application mainly depends on the implementation of the actors and of the schedulers of its actor spaces. Another important factor that influences its execution is the implementation of the exchange of messages between both local and remote actors. However, a combination of such implementations, that maximizes the quality of execution of an application, could be a bad configuration for another type of application. Moreover, different instances of the same application can work in different conditions (e.g., different number of users to serve, different amount of data to process) and so they may require different configurations.

A. Actor and Scheduler

As introduced in a previous section, actor implementations can be divided in two classes that allow to an actor either to have its own thread (from here named active actors) or to share a single thread with the other actors of the actor space (from here named passive actors). The use of active actors has the advantage of delegating the scheduling to the JVM with the advantage of guaranteeing actors to have a fair access to the computational resources of the actor space. However, this solution suffers from high memory consumption and context-switching overhead and so can be used in actor spaces with a limited number of actors. Therefore, when the number of actors in an actor space is high, the best solution is the use of passive actors whose execution is managed by a scheduler provided by the CoDE framework. Such a scheduler uses a simple not preemptive round-robin scheduling algorithm and so the implementation of the passive actor has the duty of guaranteeing a fair access to the computational resources of the actor space, for example, by limiting the number of messages that an actor can process in a single execution cycle. Moreover, is some particular applications is not possible to distribute in equal parts the tasks among the actors of an actor space and so there are some actors that should have a priority on the access to the computational resources of the actor space. Often in this situation, a good solution is the combination of active and passive actors.

B. Communication

In an actor-based system where the computation is mainly based on the exchange and processing of messages, the efficiency of the communication supports are a key parameter for the quality of applications. In CoDE both local and remote communication can be provided by replaceable components. In particular, the current implementation of the software framework supports the communication among the actor spaces through four kinds of connector that respectively use ActiveMQ [9], Java RMI [10], MINA [11] and ZeroMQ [12]. Moreover, when in an application the large part of communication is based on broadcast and multicast messages, CoDE allows to the replace the traditional individual mailbox with a mailbox that transparent extract the messages for its actor from a single queue shared with all the other actors of the actor space.

C. Experimentation

The performances of the different types of execution managers and scheduling actors can be analyzed by comparing the execution times of three simple applications on a laptop with an Intel Core 2 - 2.90GHz processor, 16 GB RAM, Windows 8 OS and Java 7 with 4 GB heap size. These examples involves four kinds of configuration:

- active, i.e., all the actor of the actor space are active actors;
- passive, i.e., all the actor of the actor space are passive actors that process all the previously received messages in each cycle of execution assigned by the scheduler;
- shared, i.e., all the actor of the actor space are passive actors and the communication among them is managed through a single queue of messages shared by the mailboxes of the actors. Actors process all the messages received in the previous scheduling cycle;
- hybrid, i.e., the actor space contains both active and passive actors. Passive actors process all the previously received messages in each cycle of execution assigned by the scheduler.

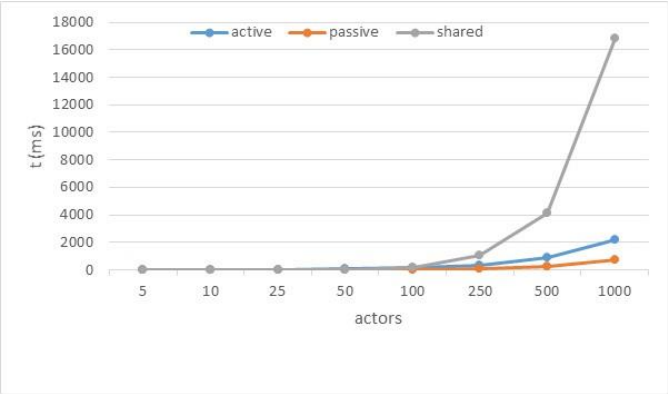


Fig. 3 Messaging example performance

The first application is based on the point-to-point exchange of messages between the actors of an actor space. The application starts an actor that creates a certain number of actors, sends 1000 messages to each of them and then waits for their answers. Fig 3 shows the execution time of the application from 5 to 1.000 actors and the best performances are obtained with the passive configuration when the number of actors increases.

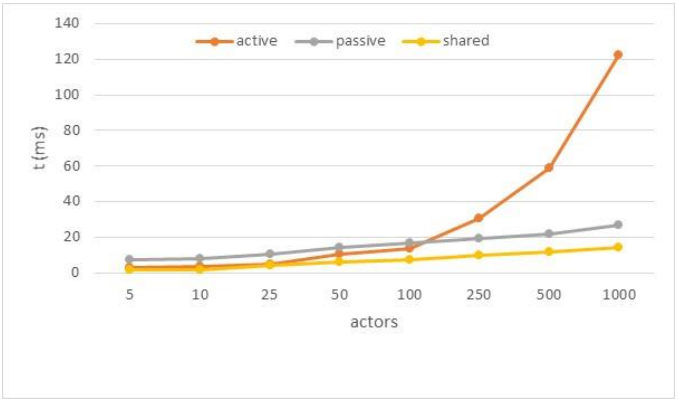


Fig. 4 Broadcasting example performance

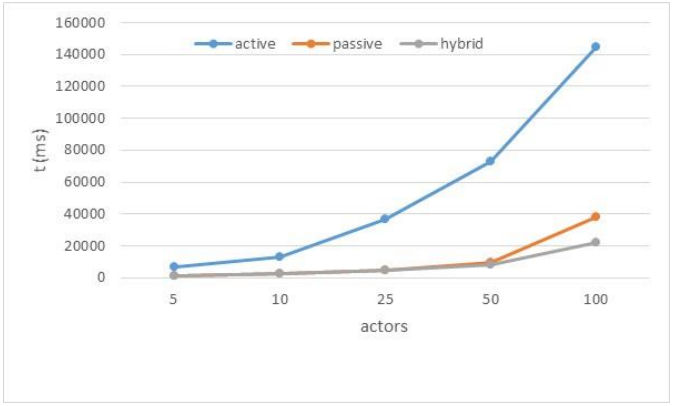


Fig. 5 Publish-subscribe example performance

The second application is based on the broadcasting of messages to the actors of an actor space. The application starts an actor that creates a certain number of actors and then sends a broadcast message. Each actor receives the broadcast message, then, in its response, sends another broadcast message and finally waits for all the broadcast messages. Fig. 4 shows the execution time of the application from 5 to 1.000 actors and the best performances are obtained with the shared configuration.

Finally, the third is a typical publish – subscribe application. In particular, there is a set of subscribers, which register their interest on the messages sent by a set of publishers. Each publisher cyclically sends a message until it reach a predefined number of messages. Each subscriber processes all the messages sent by the publishers and then kills itself. Fig. 5 shows the execution time of the application from 5 subscribers and 1000 publishers to 100 subscribers and 1000 publishers. The hybrid configuration runs the subscribers as active actors and the publishers as passive actors. The performances of passive and hybrid configurations are similar up to 250 subscribers that the best solution is the use of the hybrid and configuration.

V. Conclusion

This paper presented a software framework, called CoDE, which allows the development of efficient large actor based systems by combining the possibility to use different implementations of the components driving the execution of actors with the delegation of the management of the reception of messages to the execution environment.

CoDE is implemented by using the Java language and is an evolution of HDS [13] and ASIDE [14] from which it derives the concise actor model, and takes advantages of some implementation solutions used in JADE [15]. CoDE shares with Kilim [3], Scala [4] and Jetlang [6] the possibility to build applications that scale applications to a massive number of actors, but without the need of introducing new constructs that make complex the writing of actor based programs. Moreover, CoDE has been designed for the development of distributed applications while the previous three actor based software were designed for applications running inside multi-core computers. In fact, the use of structured messages and message patterns makes possible the implementation of complex interactions in a distributed application because a message contains all the information for delivery it to the destination and then for building and sending a reply. Moreover, a message pattern filters the input messages on all the information contained in the message and not only on its content.

Current research activities are dedicated to extend the software framework to offer it as means for the development of multi-agent systems. Future research activities will be dedicated to the extension of the functionalities provided by the software framework and to its experimentation in different application fields. Regarding the extension of the software framework, current activities have the goal of providing a passive threading solution that fully take advantage of the features of multi-core processors, of enabling the

interoperability with Web services and legacy systems [16], and of enhancing the definition of the content exchanged by actors with semantic Web technologies [17]. Moreover, future activities will be dedicated to the provision of a trust management infrastructure to support the interaction between actor spaces of different organizations [18][19]. Current experimentation of the software framework is performed in the field of the modeling and simulation of social networks [20], but in the next future will be extended to the collaborative work services [21] and to the agent-based systems for the management of information in pervasive environments [22].

REFERENCES

[1] G.A. Agha, Actors: A Model of Concurrent Computation in Distributed Systems, Cambridge, MA: MIT Press, 1986.

[2] C. Varela, and G.A. Agha, "Programming dynamically reconfigurable open systems with SALSA," SIGPLAN Notices, vol. 36, no 12, pp. 20-34, 2001.

[3] S. Srinivasan, and A. Mycroft, "Kilim: Isolation-typed actors for Java," In ECOOP 2008 – Object-Oriented Programming, Berlin , Germany: Springer, 2008, pp. 104-128.

[4] P. Haller, and M. Odersky, "Scala Actors: unifying thread-based and event-based programming," Theoretical Computer Science, vol. 410, no.2-3, pp. 202–220, 2009.

[5] Typesafe, Akka software Web site. Available: http://akka.io.

[6] M. Rettig, Jetlang software Web site. Available: http://code.google.com/p/jetlang/.

[7] J. Dedecker, T. Van Cutsem, S. Mostinckx, T. D'Hondt and W. De Meuter, "Ambient-oriented programming in ambienttalk," in ECOOP 2006 – Object-Oriented Programming, Berlin, Germany: Springer, 2006, pp. 230-254.

[8] M. S. Miller, E. D. Tribble and J. Shapiro, "Concurrency among strangers," in Trustworthy Global Computing, Berlin, Germany: Springer, 2005, pp. 195-229.

[9] B. Snyder, D. Bosnanac and R. Davies, ActiveMQ in action, Westampton, NJ, USA: Manning, 2001.

[10] E. Pitt and K. McNiff, Java.rmi: the Remote Method Invocation Guide. Boston, MA, USA: Addison-Wesley, 2001.

[11] Apache Software Foundation. (2014). Apache Mina Framework [Online]. Available: http://mina.apache.org

[12] P. Hintjens, ZeroMQ: Messaging for Many Applications, Sebastopol, CA: O'Reilly, 2013.

[13] A. Poggi, "HDS: a Software Framework for the Realization of Pervasive Applications," WSEAS Trans. on Computers, vol. 10, no. 9, pp. 1149-1159, 2010.

[14] A. Poggi, "ASiDE - A Software Framework for Complex and Distributed Systems," in 16th WSEAS International Conference on Computers, Kos, Greece, 2012, pp. 353-358.

[15] A. Poggi, M. Tomaiuolo and P. Turci, "Extending JADE for agent grid applications," in 13th IEEE Int. Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE 2004), Modena, Italy, 2004, pp. 352-357.

[16] A. Poggi, M. Tomaiuolo and P. Turci, "An Agent-Based Service Oriented Architecture," in WOA 2007, Genova, Italy, 2007, pp. 157-165.

[17] A. Poggi, "Developing ontology based applications with O3L," WSEAS Trans. on Computers, vol. 8 no. 8, pp. 1286-1295, 2009.

[18] A. Poggi, M. Tomaiuolo and G. Vitaglione, "A Security Infrastructure for Trust Management in Multi-agent Systems," in Trusting Agents for Trusting Electronic Societies, Theory and Applications in HCI and E-Commerce, LNCS, vol. 3577, R. Falcone, S. Barber, and M. P. Singh, Eds. Berlin, Germany: Springer, 2005, pp. 162-179.

[19] M. Tomaiuolo, "dDelega: Trust Management for Web Services," Int. J. of Information Security and Privacy, vol. 7, no. 3, pp. 53-67, 2013.

[20] F. Bergenti, E. Franchi and A. Poggi, "Selected models for agent-based simulation of social networks," in 3rd Symposium on Social Networks and Multiagent Systems (SNAMAS'11), York, UK: Society for the Study of Artificial Intelligence and the Simulation of Behaviour, 2011. pp. 27-32.

[21] F. Bergenti, A., Poggi and M. Somacher, "A collaborative platform for fixed and mobile networks," Communications of the ACM, vol. 45, no. 11, pp. 39-44, 2002.

[22] F. Bergenti and A. Poggi, "Ubiquitous Information Agents," Int. J. on Cooperative Information Systems, vol. 11, no. 3-4, pp. 231-244, 2002.

Agostino Poggi (M'76–SM'81–F'87) is full professor of Computer Engineering at the Department of Information Engineering of the University of Parma. His research focuses on agent, Web and object-oriented technologies and their use to develop distributed and complex systems.

He is author of more than two hundreds of technical papers in international scientific journals, books and refereed conference proceedings, and his scientific contribution has been recognized through the "System Research Foundation Outstanding Scholarly Contribution Award" in 1988 and through the "Innovation System Award" in 2001. He is currently in the editorial board of the following scientific journals: International Journal of Agent-Oriented Software Engineering, International Journal on Advances in Intelligent Systems, International Journal of Hybrid Intelligent Systems, International Journal of Multiagent and Grid Systems and Software Practice & Experience.

A Fuzzy Ontology-based Term Weighting Algorithm for Research Papers

Zeinab E. Attia

Abstract—The paper proposes a term weighting algorithm for research papers. It weights a paper's annotated keywords according to a certain view using a predefined multi-view fuzzy ontology. It solves some problems that face the current information retrieval systems which are low in precision, low in recall, inaccurate ranking the resulted documents and the multi-field topics problems. The proposed algorithm is tested and results are compared with both an ontology based term weighting algorithm and the TF-IDF algorithm. The tests show that it enhances the resulted weights accuracy.

***Keywords*—**Term weighting, Fuzzy ontology.

I. Introduction

An information retrieval system (IR) consists of a set of documents, a user query, a retrieval engine and a ranking module. It stores and indexes documents, such that when users express their information needs in a query, the retrieval engine retrieves a set of relevant documents associating a score to each one. The higher the score is, the greater the document relevance. Then the ranking module ranks them and displays them to the user [4]. In order to index a set of documents an annotation algorithm should be used.

Annotation is a process of adding metadata to certain document in order to describe it. This metadata may include a string of weighted keywords, authors' names, the publishing conference or journal, date of publishing …etc. The weight of each keyword reflects to what extend does it represents that document. The process of weighting a document's keywords is known as Term Weighting (TW).

The paper presents a domain specific Term Weighting approach based on a multi view fuzzy ontology. The paper is organized as follows: the next section presents information retrieval systems. Term weighting techniques is presented in section 3. Section 4 conducts a survey about fuzzy ontology. Some related work is presented in section 5. Section 6 presents the proposed term weighting algorithm. The proposed algorithm is tested in Section 7.

II. Information Retrieval Systems

The current Information Retrieval (IR) systems suffer from several problems. Some of them are low in recall, low in precision, inaccurate ranking for the resulted documents and inability for handling multi-field topics problem.

Recall is the proportional of the correctly retrieved documents among the pertinent documents in the collection [14]. Precision is the proportion of the correctly retrieved documents among the documents retrieved by the system [14].

Multi-field topics are topics that combine two or more fields together such as the bioinformatics that combines the medical field with the computer science field. When certain medical user searches for a bioinformatics paper, the IR system will return the same set of documents that are returned to a computer science user. So these systems do not have the ability to distinguish between results of such topics respecting the field point of view.

One of the main reasons that cause these problems is the use of inaccurate term weighting algorithms.

III. Term Weighting

Term weighting (TW) is a process to calculate a weight for each term representing a certain document. This weight reflects to what extent does this term represent that document. Due to its importance, term weighting is used in many fields such as document clustering, information retrieval (IR), and many more. Regarding IR, it enhances the recall and the precision measure. Also it enhances the rank of the retrieved documents [1].

There are several algorithms implementing the term weighting concept. These algorithms are applied either on a domain specific or on a general one. Almost all general term weighting algorithms are statistical algorithms. One of the most popular term weighting algorithms for the general Term Weighting approach is the Term Frequency Inverse Document Frequency, TFIDF. TFIDF is a statistical based method which uses equation1 [2]:

$$TFIDF_{ij}=TF_{ij}*IDF_i \quad (1)$$

$$TF_{ij}=\frac{the\ number\ of\ occurening\ the\ term\ Ti\ in\ the\ document\ dj}{the\ number\ of\ all\ terms\ in\ the\ document\ dj} \quad (2)$$

$$IDF_i=\log\frac{N}{n} \quad (3)$$

Where, TF_{ij} is the frequency of occurring the term t_i in the document d_j with respect to the number of all words in that document. IDF is the inverse document frequency. N is the number of documents in the collection. n is the number of documents that contain the term t_i.

Domain specific term weighting algorithms use domain ontology to expand a certain document keywords with their synonyms to increase their weights accuracy. These weights are calculated using some statistical formulas.

IV. Fuzzy Ontology

Ontology is "the conceptualization of a domain into a human understandable, machine readable format consisting of entities, attributes, relationships, and axioms". It is used as a standard knowledge representation for the semantic web [7].

Unfortunately, the conceptual formalism, supported by typical ontology, may not be sufficient to represent uncertain information commonly found in many application domains.

This is due to the lack of clear-cut boundaries between concepts of the domain, i.e. a concept may be a synonym for another one with a specific matching degree. Moreover, fuzzy knowledge plays an important role in many domains that face a huge amount of imprecise and vague knowledge and information, such as text mining, multimedia information system, medical informatics, machine learning, and human natural language processing. To handle uncertainty of information and knowledge, one possible solution is to incorporate fuzzy set theory into ontology [3], which leads to the birth of fuzzy ontology.

Although fuzzy ontology is used in many applications such as semantic web, multi-agent systems, information retrieval, and many more, no standard components for fuzzy ontology are found in the literature [5]. Researchers define fuzzy ontology components according to their used application and domain.

Some of such definitions are as follows:

- Fuzzy ontology is a quintuple(C, R, H, P, A), where C is a set of fuzzy concepts, R is a set of fuzzy relations between individuals, H, is a concept hierarchy, P is a set of non-taxonomic fuzzy relations between concepts, and A is a set of fuzzy axioms [6].
- Fuzzy ontology is a triple(C, I, R), where, C is a set of concepts, I is a set of individuals, and R is a set of binary relations between some elements of C and I, including two special types of fuzzy relations [8].
- Fuzzy ontology is a quadruple(C, R, P, I), where C is a set of fuzzy concepts, R is a set of binary relations, P is a set of fuzzy properties of concepts, and I is a set of individuals [10].
- Multi-Views Fuzzy Related Ontologies, MVFRO, is a couple of ($f\text{-}o_s$, $f\text{-}o_I$), where $f\text{-}o_s$ is a fuzzy ontology structure, while $f\text{-}o_I$ is the fuzzy ontology individual of concepts and relationships associated with the fuzzy ontology structure. F-os is a quintuple (C, CR, P, T, A). C is a set of fuzzy concepts. CR is a fuzzy relation between concepts. It can have more than one value each represents a certain view. P is a set of concept properties. T is a set of terms that express the concept c. A is a set of axioms [13].

V. Related Work

Two annotation techniques based on crisp ontology are proposed in [9]. Each technique can annotate a set of documents with a string of weighted keywords in two steps.

The first step is to annotate documents with a string of keywords. This string is then entered the second step to be weighted. These weighted keywords are then stored in a relational database such that each tuple in it indicates that a document d_i is indexed by a term t_k with a weight w_j. The first annotation technique uses an NLP annotation algorithm to annotate a certain document with a string of keywords. Then, these keywords are weighted using an adapted TF-IDF algorithm. This adapted algorithm is the frequency of the occurrence of each semantic entity in the ontology or any of its associate keywords within a document. Such an algorithm takes pronoun into account. The second technique uses a contextual semantic information based algorithm to annotate a certain document with a string of keywords. Then, theses keywords are weighted using a fusion weighting algorithm.

An annotation system performing a clustering process based on a concept weight supported by crisp domain ontology is proposed in [11]. The system is divided into three major modules; document preprocessing, calculating a concept weight based on ontology, and clustering documents with the concept based. The weighting module is calculated through equation 4 [11]:

$$W = Len \times Frequency \times Correlation\ Coefficient + Probability\ of\ concept \quad (4)$$

Where, W is the weight of a certain keyword. Len is the length of that keyword. Frequency is times which the words appear, and if the concept is in the ontology, then correlation coefficient =1, else correlation coefficient=0.Probability is based on the probability of the concept in the document. The probability is estimated by equation 5 [11]:

$$P\ (concept) = \frac{Number\ of\ Occurences\ of\ the\ Concept}{Number\ of\ Occurences\ of\ All\ Concept\ in\ Document} \quad (5)$$

A new weighting method based on statistical estimation of a word importance for a particular categorization problem is proposed in [12]. This weighting also has the benefit that it makes feature selection implicit since useless features for the categorization problem considered get a very small weight.

These algorithms are ontology based term weighting algorithms. They can weight a document keyword according to only one view. So, they do not consider the multi-field topics problem. Also, all of them suffer from inaccurate weights as the result of using crisp ontology. Let's consider this example, if you have two Information Retrieval (IR) papers, one is about keyword-based IR and the other is about concept-based IR. Ofcouse, any IR expert will give the concept-based IR paper a higher weight than the keyword-based one. Using Ontology, it will give the two papers the same weight. Since ontology represents a certain domain with a set of concepts and crisp relations between them. On the other hand, Fuzzy Ontology, FO, will give the concept-based IR paper higher weight than the keyword -based IR one. So, it mimics the expert opinion. So the proposed algorithm is fuzzy ontology based term weighting algorithm.

VI. The Proposed Term Weighting Algorithm

The proposed algorithm is a semantic based term weighting algorithm. It considers the multi-field topics problem. It calculates a weight for each annotated keyword in a certain paper according to a specific field or view. This weight reflects to what extend this keyword represents that paper according to the specified view. The implemented algorithm uses a predefined multi-view fuzzy ontology and a stemmer algorithm. The algorithm aims to enhance the resulted weights accuracy in a specific view through:

- Using a multi-view (multi-field) fuzzy ontology instead of the crisp one for expanding each annotated keyword in the keyword zone respecting a certain view. This solves:
  1- The multi-field topics problem. As this helps in annotating a certain paper with such topic in the two fields.
  2- The recall, precision and the inaccurate ranking problem through using fuzzy ontology instead of crisp ones.
- Arranging the paper expanded keyword list in a descending order according to each keyword n-grams, the number of terms in each keyword,
- Replacing each pronoun with its referred noun, instead of removing it as stop word,
- For titled sections, they are annotated with their titles not with keywords listed in the paper keyword zone,
- Keywords written with different style (Bold, Italic, Underlined) will have higher weights,
- Keywords written as a section title or as a figure caption also will have a higher weight,
- For well written papers, the proposed algorithm can be applied on each paragraph main sentence as it reflects its main idea, instead of working on the whole paper. This will reduce the execution time.

A. The Proposed Term Weighting Algorithm Phases

The proposed algorithm phases are as follow:

1) Preprocessing phase

Firstly, the paper is divided into different weighted zones. Zone is one of the standard sections in any research paper, e.g., Title, Abstract, Introduction, Related Work, References, etc. The weight of each zone reflects the role of this zone in the paper, e.g., {(title, 1), (introduction, 0.5)……}.
Secondly, remove the reference zone from the paper. Then, extract the annotated keywords from the keyword zone associating each of them with a weight, w

$$PKS= \{(k_{i0}, w_{k_{i0}})\},\ i=1....n,$$

Where PKS is the Paper Keyword Set. k_{i0} is a keyword in the keyword zone. n is the number of keywords in the keyword zone. W_{ki0}reflects to what extent k_{i0} represents the paper. $W_{ki0} \in]0, 1]$, its value is calculated through this algorithm.

2) Annotating the paper zones

This phase is responsible for annotating the paper zones. Each zone in the paper is annotated with the PKS and its expansion set. Each keyword in the PKS is expanded with all concepts and all terms related to it with a certain threshold in specific view using the predefined multi-view fuzzy ontology. Each keyword is stemmed and then expanded through the following steps:

- If it is represented in the fuzzy ontology as a concept, expand it with all its related concepts, and terms that can represent it with degree greater than or equals to a certain threshold in the considered view.
- If it is represented in the fuzzy ontology as a term, expand it with all its related terms, and concepts that it can represent in the document domain with a degree greater than or equals to a certain threshold in the considered view.

So that:

$$PKS= \{(k_{i0}, w_{k_{i0}}), (k_{ij}, w_{k_{ij}})\},\ i= 1...n,\ j=1.....m,$$

where, k_{i0} is a keyword in the keyword zone. w_{ki0} reflects to what extent k_{i0} represents the paper, $w_{ki0} \in]0, 1]$, its value is calculated through this algorithm. n is the number of keywords in the keyword zone. k_{ij} is an expanded keyword for the keyword k_{i0}. m is the number of the k_{i0}'s expanded keywords. w_{kij} reflects to what extent k_{ij} is related to k_{i0}, $w_{kij} \in [0, 1]$. After expanding all the paper keywords, arrange the PKS in a descending order according to the number of n-grams of each keyword in it.

3) Annotating the paper sections

This phase determines the paper sections and annotates each of them with its title. A section is a section entitled with one of the keywords in the PKS; otherwise it is treated as a zone. "Fuzzy Ontology" and "Term Weighting" are section example in this paper. To annotate a certain section, consider its title to apply the following steps on it:

1. Return each pronoun to its referred noun.
2. Remove all stop words.
3. Stem the remaining keywords.
4. If this title includes one of the keywords in PKS, then
   i. Put these keywords in a Section Keyword Set, SKS, associating each of them with a weight, w.

$$SKS= \{(s_l k_{i0}, w_{s_l k_{i0}})\},\ i=1....n,\ l=1...p$$

where s_{lki0} is a keyword extracted from the section l title. n is the number of keywords that are extracted from this section's

title. p is the number of sections in the paper. w_{slkio} reflects to what extent the keyword k_{i0} represents the section l, $w_{slkio} \in [0, 1]$, its value is calculated through executing the algorithm.

ii. Expand each keyword in the SKS in the given view using the predefined domain fuzzy ontology using the same methodology described the second phase, so that,

$SKS= \{(s_l k_{i0}, w_{s_l k_{i0}}), (s_l k_{ij}, w_{s_l k_{ij}})\}$, i= 1…n, j=1…..m, l=1…p

where, s_{lki0} is a keyword extracted from the section s_l title. w_{slkio} reflects to what extent the keyword k_{i0} represents the section s_l, $w_{slki0} \in [0, 1]$, its value is calculated through this algorithm. p is the number of sections in the paper. m is the number of k_{i0}'s expanded keywords. s_{lkij} is an expanded keyword for the keyword k_{i0}. w_{slkij} reflects to what extent kij is related to k_{i0}.

iii. Arrange the SKS list in a descending order according to the number of n-grams of each element belongs to the SKS.

4) Weighting phase

This phase calculates the weight of each keyword in the keyword zone using equation 6 as the summation of its weight in each zone and in each section.

$$w_{k_{io}} = \sum_{l=1}^{y} w_{z_l k_{io}} + \sum_{l=1}^{p} w_{s_l k_{io}} \quad (6)$$

Where w_{zlkio} is the weight of the keyword k_i in zone l. w_{slki0} is the weight of the keyword k_i in section l. p is the number of sections in this paper. Y is the number of zones in this paper.

To calculate the weight of a certain keyword in a zone or in a section, the proposed algorithm:

1. For only well written papers, considers the main sentence of each paragraph in this zone or in this section, otherwise, consider the whole paper paragraphs.
2. Returns each pronoun in it to its referred noun.
3. Removes all stop words.
4. Stems each of the remaining keyword.
5. Calculates the weight of the keyword ki that is belongs to PKS in a certain zone as in equation 7.

$$w_{z_l k_{io}} = w_{z_l} * (freq_{z_l k_{io}} + \sum_{j=1}^{m} w_{k_{ij}} * freq_{z_l k_{ij}}) \quad (7)$$

$$freq_{z_l k_{io}} = \frac{\text{number of occurring } k_{io} \text{ in zone } z_l}{\text{number of words in the same zone}} \quad (8)$$

$$freq_{z_l k_{ij}} = \frac{\text{number of occurring } k_{ij} \text{ in zone } z_l}{\text{number of words in the same zone}} \quad (9)$$

Where w_{zlki0} is the k_{i0}'s weight in the zone z_l. w_{zl} is the zone z_l's weight. $freq_{zlki0}$ is the frequency of occurring the keyword k_{i0} in the zone z_l with respect to the number of words in the same zone. $freq_{zlkij}$ is the frequency of occurring the expanded keyword k_{ij} in the zone z_l with respect to the number of words in the same zone. w_{kij} is the weight that reflects to what extend the expanded keyword k_{ij} is related to the keyword k_{i0}.

Any keyword in the paper can matches with at most one element in the PKS, e.g., consider a PKS= {fuzzy ontology, fuzzy, ontology} and a sentence "fuzzy ontology represents a fuzzy domain" after processing the sentence will be "fuzzy ontology represent fuzzy domain". The number of occurring the word fuzzy ontology in this sentence is 1, zero for the word ontology, and only 1 for the word fuzzy.

In the same manner, the weight of the keyword ki in section l is calculated. A section weight is calculated using equation 10 as the ratio between its number of words to the paper number of words.

$$w_{s_l} = \frac{number\ of\ words\ in\ section\ s_l}{number\ of\ words\ in\ the\ document} \quad (10)$$

B. The proposed Algorithm

The algorithm is illustrated as in algorithm1.

Algorithm 1: Term weighting of research paper in certain View

Input: research paper in a certain view, a predefined multi-view fuzzy ontology

Output: research paper annotated with a set of weighted keywords in the specified view

Steps:

1. Divide the paper into different weighted zones
2. PKS= expand all keywords in the keyword zone according to the given view using the predefined fuzzy ontology
3. Arrange all PKS in decreasing order according to each keyword n-gram
4. annotate each zone with the PKS
5. for each zone, calculate the weight of each keyword in PKS
6. **For** each section
7. annotate it with its title
8. SKS= expand this annotation using the predefined multi-view fuzzy ontology
9. arrange SKS in a descending order with respect to each keyword n-gram
10. calculate the weight of each keyword in SKS
11. **End for**
12. Calculate the weight of each keyword in the keyword zone through summing its value from each zone and each section.

After applying the proposed term weighting algorithm, the resulted weighted keywords are stored in a relational database. As shown in figure 1, this database has three tables; *documents*, *fuzzyDocumentClasses*, and *fuzzyDocumentTerms* tables. *Documents* table stores all information needed to access any document such as, document name, and document path. Also, it gives each one a unique identifier. The annotated weighted keywords of a certain document are stored in the other two tables; *fuzzyDocumentClasses* and *fuzzyDocumentTerms*. If these keywords are represented as concepts in the used fuzzy ontology, then this annotation is stored in *fuzzyDocumentClasses* table. This table is connected with the fuzzy ontology database through the foreign key c_id (concept_id). While, if these keywords are represented as terms in the used fuzzy ontology then, these annotation type is stored in *fuzzyDocumentTerms* table. This table is also connected with the fuzzy ontology database through the foreign key t_id (term_id).

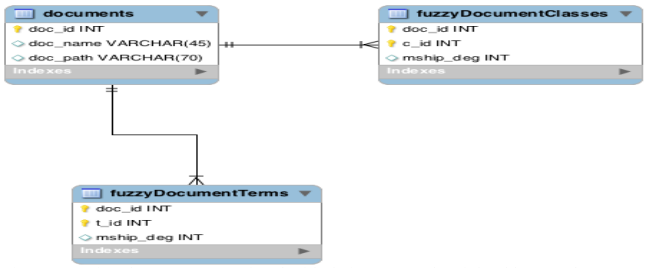


Figure1: storing documents associated with their annotated weighted keyword

VII. Tests and Results

To test the proposed term weighting algorithm and compare it with the TF-IDF and Fernández algorithms, a small document collection on a specific domain of interest is assembled. The collection is composed of 100 research paper on computational Intelligence. These papers are classified into 20 papers about Fuzzy ontology information retrieval, 20 about Fuzzy information retrieval, 20 about Ontology based information retrieval, 20 about ontology and 20 about Fuzzy ontology.

Figure2 presents a comparison between the three term weighting algorithms according to ranking a set of document respecting a specific topic. It is obvious from that the figure that:

- IF-IDF has a low ranking accuracy as it depends on counting the frequency of a keyword in the document without considering its synonyms or its related keywords.
- Fernandez algorithm enhances the ranking accuracy with respect to TF-IDF for Ontology based IR, Fuzzy IR, Fuzzy Ontology and Ontology. This is due to using ontology for expanding a certain keyword with its synonyms and its related keywords.
- Fernandez algorithm decrease the ranking accuracy for the Fuzzy Ontology based IR. As the result of the used general ontology that expand a certain keyword with its related keywords in all domains.
- The proposed algorithm enhances the ranking accuracy with respect to TFIDF and Fernandez algorithms. This is due to expanding a certain keyword using a specific domain fuzzy ontology, replacing each pronoun to its related noun, arranging each annotated keyword in a descending order respecting its n-gram, annotating each section with its title, dividing the paper into different weighted zones.

From this collection, 20 research papers are well written.

Figure 3 shows the resulted weights for a sample of eight papers from those 20 ones. Each chart in this figure represents the weights for each keyword in a specified paper after applying the proposed algorithm on its paragraphs sentences gradually (the main sentence only, the first two sentences, the first three sentences ….. and on all sentences. Considering figure 3, it is obvious that:

- Applying the proposed algorithm into only the main sentence in each paragraph, approximately, gains the same relative weights for the papers key words. Accordingly, the time of annotation operation will be reduced.
- After the second line of applying the proposed algorithm the weight of a certain keyword becomes approximately stable. Consequently, with well written papers, applying the proposed algorithm to the two first sentences or even to just the main sentence in each paragraph becomes satisfactory.

Table1 shows that working on a well written paper main sentence instead of working on all of its sentences decreases the execution time while the relative ranking accuracy remains the same.

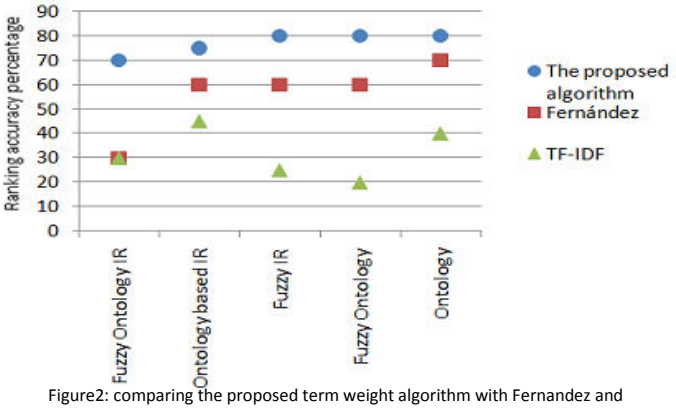


Figure2: comparing the proposed term weight algorithm with Fernandez and the TF-IDF according to ranking accuracy percentage

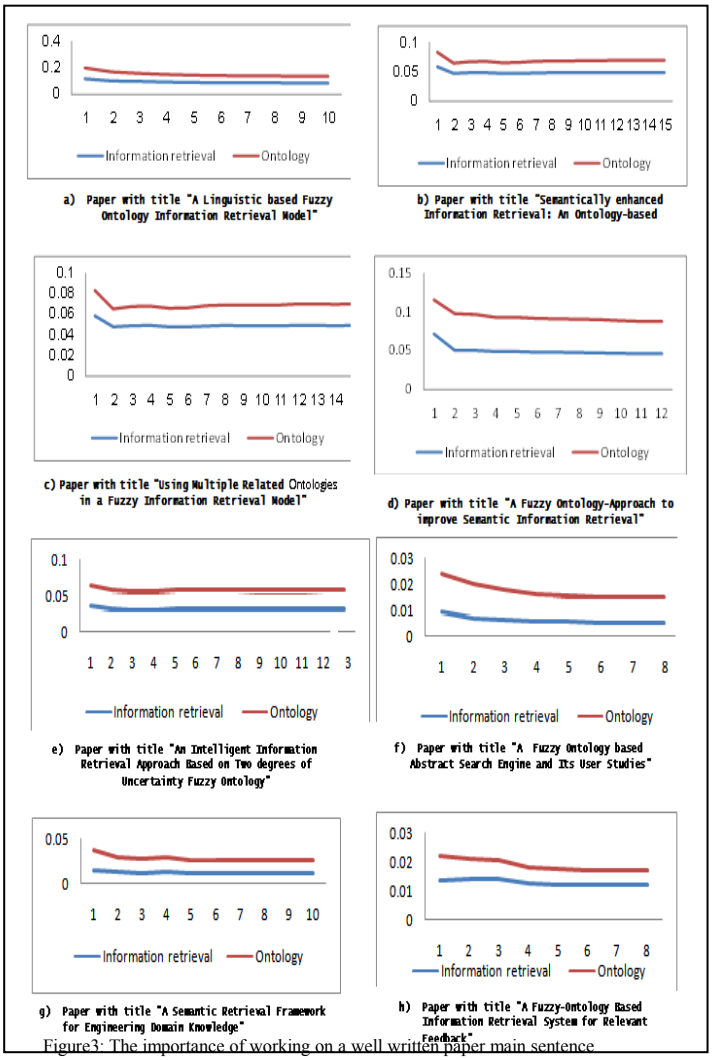


Figure3: The importance of working on a well written paper main sentence

Table 1: Ranking a set of 10 research papers after applying the proposed term weighting algorithm on all lines and on just both of main and subordinate sentences in each paragraph

Paper title	Applying the proposed algorithm on all lines		Applying the proposed algorithm on 2 lines	
	Keyword weights	Elapsed time (SEC)	Keyword weights	Elapsed time (SEC)
Semantically enhanced Information Retrieval: An ontology-based approach	0.051360	6 sec	0.04276	5.6 sec
A Linguistic-based Fuzzy Ontology Information Retrieval Model	0.039556	2 sec	0.03997	1.5 sec
A Semantic Retrieval Framework for Engineering Domain Knowledge	0.039420	0.8 sec	0.02851	0.5 sec
Using Multiple Related Ontologies in a Fuzzy Information Retrieval Model	0.034245	1.5 sec	0.02448	1 sec
An Intelligent Information Retrieval Approach Based on Two Degrees of Uncertainty Fuzzy Ontology	0.030659	0.5 sec	0.02255	0.4 sec
A Fuzzy-Ontology Based Information Retrieval System for Relevant Feedback	0.018974	2.6 sec	0.01970	2 sec
A Fuzzy Ontology-Approach to improve Semantic Information Retrieval	0.017925	1.7 sec	0.01823	0.7 sec
A Fuzzy Ontology based Abstract search engine and its users studies	0.017908	0.7 sec	0.01781	0.6 sec

VIII. Conclusion

Compared with TF-IDF and Fernández algorithms, the proposed algorithm enhances the tested documents weights accuracy. This is due to dividing the paper into different weighted zones, using fuzzy ontology instead of using crisp one, arranging the expanded list in a descending order respecting the number of n-grams of each keyword in it, annotating each paper section with its title, returning each pronoun to its referred noun.

For well written papers, applying the proposed algorithm on their paragraphs main sentences instead of working on the whole paper decreases their time of execution while the ranking accuracy remains unchanged.

References

[1] S. Klink, K. Kise, A. Dengel, M. Junker, and S. Agne, "Document Information Retrieval," Digital Document Processing, Springer-Verlag London Limited 2007.

[2] Salton, Gerard, Buckley, and Chris, "Term weighting Approaches in Automatic Text Retrieval," Information Proceeding and Management Vol.32 (4), pp. 431-443, 1996.

[3] J. Zhai, Y. Liang, Y. Yu and J. Jiang "Semantic Information Retrieval Based on Fuzzy Ontology for Electronic Commerce," JOURNAL OF SOFTWARE, VOL. 3, NO. 9, DECEMBER 2008.

[4] M. A. A. Leite and I. L. M. Ricarte, "Relating ontologies with a fuzzy information model," Journal Of Knowledge and Information System, pp. 619-651, 2013.

[5] F. B. Ortenga, M. D. Calvo-Flores, "Managing Vagueness in Ontologies," PHD Dissertation, Granada, October 2008.

[6] E. Sanchez, T. Yamanoi, "Fuzzy Ontologies for the Semantic Web," the 7th International Conference on Flexible Query Answering Systems (FQAS 2006), Vol. 4027, pp.691-699, 2006.

[7] Q. T. Tho, S. C. Hui, A. C. M. Fong, T. H. Cao," Automatic Fuzzy Ontology Generation for Semantic Web," IEEE transaction on knowledge and data engineering, Vol. 18, No.6, June 2006.

[8] Y. Ling, H. M. Gu, X. W, J. Q. Shi, "A Fuzzy Ontology and Its Application to Computer Games," the 4th International Conference on Fuzzy Systems and knowledge Discovery (FSKD 2007), Vol. 4, pp.442-226, 2007.

[9] M. Fernández, I. Cantador , V. López, D. Vallet, P. Castells, E. Motta,"Semantically enhanced Information Retrieval: An ontology-based approach," Web Semantics: Science, Services and Agents on the World Wide Web 9, pp. 432-452, 2011.

[10] Y. Cai, H. F. Leung, "A Formal Model of Fuzzy Ontology with Property Hierarchy and Object Membership," the 27th International Conference on Conceptual Modeling (ER 2008), Vol. 5231, pp.69-82, 2008.

[11] H. H. Tar, T. T. S. Nyunt, "Ontology-Based Concept Weighting for Text Documents," International Conference on Information Communication and Management, vol.16, 2011.

[12] P. Soucy, G. W. Mineau, "Beyond TFIDF Weighting for Text Categorization in the Vector Space Model," The 19th International Joint Conference on Artificial Intelligence, pp. 1130-1135, 2005.

[13] Z. E. Alarab, A. M. Gadallah, H. A. Hefny, "An Enhanced Model For Linguistic-based fuzzy ontology," the 47th Annual Conference on Statistics, computer sciences and operation research, pp. 49-62, 2012.

[14] C. D. Manning, P. Raghavan, H. Schütze, "Introduction to Information retrieval: Evaluation in information retrieval," Cambridge University Press, 2008.

Barriers to the development of cloud computing adoption and usage in SMEs in Poland

Dorota Jelonek, Elżbieta Wysłocka

***Abstract*—** Entrepreneurs who consider changing the traditional IT management model to cloud computing have concerns about the security and confidentiality of data, the integration of internal IT infrastructure with new solutions, insufficient availability of services through the Internet or the quality and efficiency of services. The article confirmed the hypothesis that Polish SMEs managers perceive barriers to implementation and development of cloud computing in a similar manner to managers from other EU countries. In addition, it has been shown that mental barriers, including lack of knowledge and lack of trust are the most serious barriers to the dissemination of the cloud computing model. The biggest legal barriers are related to concern about data security, in particular the protection of personal data, whereas among technical barriers slow Internet connection" is not a significant obstacle in the development of cloud computing in the opinion of Polish managers. This confirms a steadily increasing broad access to high-speed Internet. The study also proved that the vast majority of Polish SMEs managers evaluate the decision of implementing cloud computing as highly risky, especially in the context of losing full control over information resources and the fear of disclosing them without the owner consent.

***Keywords*—**cloud computing, SMEs, barriers: mental, legal, technical.

I. Introduction

Cloud computing is nowadays one of the most important trends in the development of new business models, co-defining the organizational structure and the enterprise management methods. In the opinion of many managers as well as in numerous scientific publications, cloud computing improve business competitiveness by reducing costs and increasing flexibility.

Any concerns that accompany the transition to cloud computing will gradually disappear with the growing popularity of this solution and the increase of managers knowledge about it. The rising number of managers who believe that cloud computing usage is safe and may bring many tangible benefits may additionally help in the eliminations of these concerns. As every new model of IT service, cloud computing faces various barriers that hinder its implementation in enterprises. The current model of IT resources management is changed and it is followed by migration of data from own servers, therefore many managers associate it with losing control over their IT resources. What is more, in the traditional model a user can select and configure software more freely than in the cloud computing model, which imposes certain requirements and restrictions on applications and their functions.

Partial IT outsourcing is a common practice among SMEs in Poland, however a complete technology outsourcing of business processes is possible only due to development and increased accessibility to the cloud computing model.

The purpose of this article is to identify the barriers which prevent managers of SMEs sector in Poland from implementing cloud computing solutions. Results were compared with the outcomes of research carried out on the sample companies based in European Union. The authors aim to confirm the following hypothesis:

H1: Polish SMEs managers perceive barriers to implementation and development of cloud computing in a similar manner to managers from other EU countries.

H2: Mental barriers, including lack of knowledge and trust are the most serious barriers to the dissemination of the cloud computing model.

H3: The biggest legal barriers are related to the concern about data security, in particular the protection of personal data.

H4: Among the technical barriers "slow Internet connection" is not a significant obstacle in the development of cloud computing in the opinion of Polish managers. This confirms a steadily increasing broad access to high-speed Internet.

H5: The vast majority of Polish SMEs managers evaluate the decision of implementing cloud computing as highly risky, especially in the context of losing full control over information resources and the fear of disclosing them without the owner consent.

This article is organized as follows. Section 2 presents the issue of cloud computing model. Then, in Section 3, types of cloud computing services were described. Section 4 presents cloud computing from the perspective of SMEs, taking into account potential benefits and concerns of this decision. In the next section the research method was discussed, additionally the research sample and the basic questions included in the survey for the identification of mental, technical and legal barriers were presented. The results of the study were compared with the results of the research conducted on a sample companies from EU region, and the insightful analysis

ISBN: 978-1-61804-236-1

of the results confirmed the positive verification of all research hypotheses.

II. The issue of cloud computing model

The concept of cloud computing was first introduced by S.E. Gilleta i M. Kapora in 1996 [1] and formulated by them definition of this term is still valid.

Cloud computing is a relatively new business model in the computing world. It is adopted by enterprises [2] [3], public sector [4], regional business community [5] and many other organizations. According to the official NIST definition, "cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction"[6].

Armbrust et al. [7], define that, "cloud computing refers to both applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services".

Enterprises can use this solution to outsource the administration of databases, management of applications and information systems, hence they do not need to have storage for servers and other infrastructure anymore.

The cloud computing model comprises five essential characteristics. The characteristics are described as on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service [8]. On-demand self-service denotes the unilateral provisioning of resources without human interaction with the provider while broad network access means that services are delivered over a network [9]. Resource pooling is the set of resources such as storage, processing, memory, bandwidth, which can be offered to many clients. Rapid elasticity indicates that resources are dynamically scaled up and down with demand and, finally, measured service refers to the automatic control and optimization of resources through pay-per-use metering capabilities [9].

III. Types of cloud computing services

The majority of literature reviews define a Cloud Computing Framework as a Service Oriented Architecture (SOA) [10].

Depending on how advanced cloud computing is, three basic types or levels of this service are currently distinguished:

- Infrastructure as a Service (IaaS),
- Platform as a Service (PaaS),
- Software as a Service (SaaS).

Each of the service types serve different purposes and target different customers however they share a common business model that is that they 'rent' the use of their computing resources including services, applications, infrastructures, and platform to customers[11].

Infrastructure as a Service (IaaS) means using computer hardware via the Internet. IaaS is divided into Compute Clouds and Resource Clouds [10]. Compute Clouds provide access to computational resources such as CPUs, hypervisors and utilities. Resource Clouds contain managed and scalable resources as services to users.

Infrastructure as a Service model refers to the tangible physical devices (raw computing) like virtual computers, servers, storage devices, network transfer, which are physically located in one central place (data center) but they can be accessed and used over the internet using the login authentication systems and passwords from any dumb terminal or device [12]. Examples: Amazon S3 (Simple Storage Service), Amazon EC2 (Elastic Compute Cloud) and Rackspace Cloud Servers.

Platform as a Service (PaaS) is a more advanced level of cloud computing service than IaaS. PaaS provides a full or partial application development environment that enables developers to access resources for application development and to collaborate with others online.

This solution offers specific services allowing treatment of the infrastructure elements as one entity and using a single virtual supercomputer on which – thanks to special software components – scalable applications can be developed. PaaS offers an agile development environment that makes it easier for IT professionals to develop applications quickly and to adopt them instantly because it eliminates the wait for deployment of suitable hardware and software for the applications [13]. Examples: Microsoft Azure Service Platform, Saleforce - Force.com, Google App Engine and Amazon Relational Database Services.

Software as a Service (SaaS) model is the most elaborate level of cloud computing. User gets access not only to the hardware infrastructure, but also to defined IT applications. Software as a Service model can be understood as the variation to the application service provider model (ASP), where customers pay, rent, or subscribe to applications or services from the cloud providers to access applications or services such as online storage and database capabilities via the Internet [14]. A client does not bear the costs of purchasing software licenses and only pays for each usage of specific functions of programs running on a provider's server. An access to selected functions "on demand" is an advantage. SaaS solutions can provide services such as Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), Human Resource Management (HRM), finance and accounting systems, analytical systems. Some office applications and email and web pages management systems are also available. Examples: Salesforce, Netsuite and Google Apps.

Enterprises can choose how to implement the services in the cloud computing model from several types of "clouds": private clouds, public clouds, community clouds and hybrid clouds.

Private clouds is a solution designed for a specific organization and used exclusively by it. Private clouds give organizations more control over security, transparency and compliance but require substantial capital, operational

expenditures and a highly proficient IT team.

In public clouds, the infrastructure is owned by a single provider however it is addressed to the general public or a specific industry. Google Apps is a good example of this kind of cloud.

Community clouds provide cloud infrastructure for several organizations and it supports specific communities with common goals (e.g. policy, mission, security requirements). That solutions have the advantage of cost efficiency compared to private clouds.

Hybrid clouds are combinations of two or more "clouds" (private, public, or community) which are unique entities however linked by a single technology.

IV. Cloud computing from the SMEs perspective

Cloud computing is definitely making waves with SMEs and is slowly creeping into their business strategy formulation and implementation now and in the near future. Basing on the research results of Gupta, Seetharaman and Raj SMEs are not hesitant to incorporate cloud into their business strategy despite the few concerns being cited by industry pundits [12]. Decisions to implement cloud computing model are frequently preceded by identifying benefits and possible risks. The list of factors that influence decisions about the usage of cloud computing can be very long and contain factors specific to the region, industry, business model and management style. The attempt to assess the impact of five selected factors on the implementation of cloud computing in SMEs have been undertaken by P. Gupta, A. Seetharaman and J. R. Raj [12]. Firstly, ease of use and convenience is the biggest favorable factor followed by security and privacy and then comes the cost reduction. The fourth factor reliability is ignored as SMEs do not consider cloud as reliable. Lastly but not the least, SMEs do not want to use cloud for sharing and collaboration and prefer their old conventional methods for sharing and collaborating with their stakeholders [12].

Among the most frequently mentioned changes associated with introducing cloud computing model many scientist list cost savings and operational flexibility [15]. Cloud computing model reduces fixed costs associated with the purchase of infrastructure and eliminates the need of certain infrastructure (and its future software update) purchase. Hence the energy costs of infrastructure usage are decreasing and the savings in lowering stuffing costs are visible, due to the fact that there is no need to hire employees to operate their own systems.

This situation affects the enterprise flexibility both in terms of the resources volume and access to the latest technology solutions. What is more the ability of accessing data through the internet allows managers to read or edit information resources of the company from any place, at any time and by using various devices.

The main perceived benefits of the cloud computing model in SMEs were also the subject of The European Network and Information Security Agency research [16]. Fig. 1 shows the results of the study.

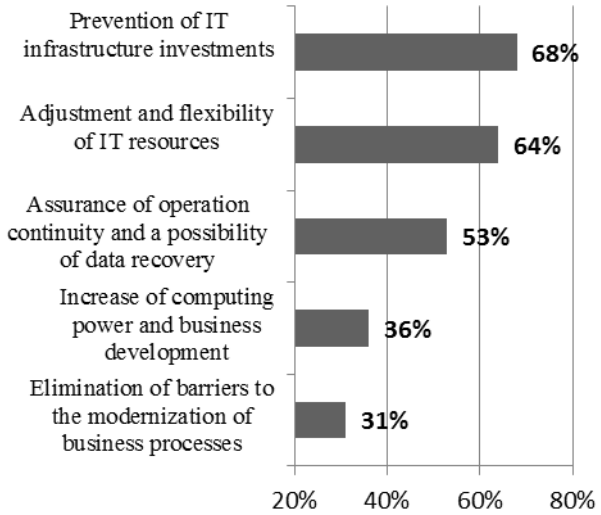


Fig. 1 Benefits of cloud computing model in SMEs [12]

Costs savings are seen as key benefits of cloud computing and 68% of SMEs indicated that this solution would help in avoidance of investments in IT infrastructure. Adjustment and flexibility of IT resources is also of a great importance for companies (64%), especially when the company do not need to worry about software updates or a new product versions or any other kind of necessary actions that has to be carried out in order to adapt systems to the changes in regulations (e.g. changes in tax rates). In the opinion of 53 % of respondents cloud computing provides business operation continuity and it allows data recovery, however the rest of the respondents (47%) are worried about the efficiency of Internet connection , the security of data transfer and its storage. From the point of view of SMEs the increase of computing power and business development (36%) and the elimination of barriers to modernization of business processes (31%) are not very important.

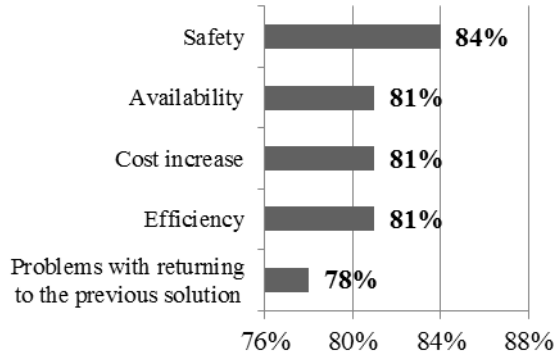


Fig. 2 Drawbacks of cloud computing model in SMEs [16]

The data presented in Fig. 2 confirm that companies have some serious concerns hindering full adoption of cloud computing model. The percentage of respondents who believe that all factors are important is very high. The biggest concerns about cloud computing relate to data security (84%). However

there are relatively small differences in answers ranging from 78% to 84% of respondents. This indicate that SMEs believe that each of the five indicated factors is a drawback or concern of cloud computing. Presented results justify the need of conducting further research on the topic of barriers to the development of cloud computing identification, in order to mitigate these barriers in the future hence dispel the fears and concerns of managers.

V. BARRIERS TO ADOPTION AND USAGE OF CLOUD COMPUTING. RESEARCH RESULTS.

The SMEs sector is dynamic and open to innovations and new solutions, however in the case of cloud computing decisions are thought through carefully, without a rush. Even the awareness of cloud computing benefits do not convince managers to its usage. The effect of "scale leveling" is the undoubted benefit of using cloud computing and Internet in business. It means that small enterprises can benefit from more advanced systems and application by incurring only the costs of actual resources used without having to bear the investment costs.

The attempt of seeking an answer to the question "What prevents SMEs based in various EU countries from implementation and further development of cloud computing?" has been undertaken in the IDC report [17]. The report was an inspiration for the authors of this study to verify the perception of barriers by the Polish SMEs.

Polish companies participated in IDC study in 2012, however the specificity of the economy of each country is different, the determinants of doing business are various, the internet infrastructure level differs, other regulations are not the same either, therefore in these contexts the perception of barriers to cloud computing implementation may be unlike those in EU.

There are 1.8 million enterprises in Poland, from which 99.8% are small and medium-sized enterprises [18]. In the light of the above, presented study results carried out on a sample of 134 companies are not representative to the entire Polish SME sector. Nevertheless in the authors' opinion, the results are an interesting attempt of building the image of Polish SMEs, which face the challenge of cloud computing implementation.

The study involved: 80 micro-enterprises, 42 small-enterprises and 10 medium-enterprises from the fallowing sectors: manufacturing - 39, trade - 65 and services – 30.

The study was conducted in January - February 2014 with the use of electronic questionnaire. Link to the survey was given via email sent to the companies. Questionnaire return rate was 27%. The results of few chosen survey questions are presented below. In order to examine the relevance of each barrier to the cloud computing implementation decision, twelve questions from IDC report [17] research carried out on sample european enterprises were used.

- Security & data protection: "We are worried about the security and data protection guaranteed by cloud services"
- Trust: "It is difficult to judge which cloud services are trustworthy"
- Data location: "We do not know and/or cannot control the location of our corporate data"
- Local support: "There is no local support for the services"
- Change control: "We cannot control software changes and upgrades made by the vendor"
- Ownership of customisation: "We do not know who owns the customisations/changes we make to the cloud services"
- Evaluation of usefulness: "We do not know how to evaluate the usefulness of cloud service for our organization"
- Slow Internet connection: "Our Internet connection(s) is/are not reliable or fast enough"
- Local language: "There is no local language version of the services"
- Tax incentives: "Tax and other incentives make buying with capital more attractive than paying for what we use on subscription"
- Legal Jurisdiction: "If we have a dispute with the cloud service provider, I may have to go to court in another country inside the EU"
- Data Access and Portability ""Concern about our ability to move data from one vendor to another or onto our own IT"

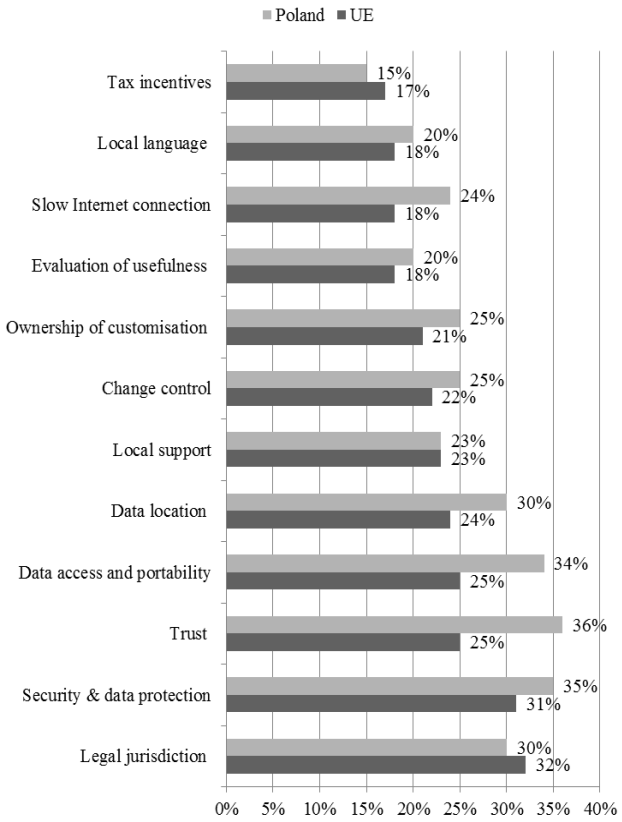


Fig. 3 The relevance of the barriers to the cloud computing implementation

Respondents rated each response on a scale: low barrier, average barrier, high barrier. The research results including only the "high barrier" response are shown in Fig. 3. For

comparison, results of research carried out for EU companies in different countries have been added.

Most serious barriers in the assessment of Polish SMEs are: "trust" (36%), "security and data protection" (35%) and "data access and portability" (34%). The lowest-rated barriers are as follows: "tax incentives" (15%) and the "local language" (20%). According to the respondents of the EU enterprises the most serious barrier are "legal jurisdiction" (32%), "security and data protection" (31%), "trust" (25%) and "data access and portability" (25%).

Had to create a ranking of the five most serious barriers, the same barriers would have been indicated by both Polish SMEs and enterprises based in EU countries (only in a different order).

Table 1. Relevance of mental, legal technical barriers to cloud computing adoption

Barriers	The relevance of barrier to cloud computing adoption			
	Low barrier	Average barrier	High barrier	
Mental barriers				
Trust	10%	54%	36%	38% of all responses was a „high barrier"
Data access and portability	14%	52%	34%	
Data location	22%	48%	30%	
Evaluation of usefulness	30%	50%	20%	
Legal barriers				
Local support	41%	36%	23%	29% of all responses was a „high barrier"
Legal jurisdiction	14%	56%	30%	
Ownership of customisation	18%	57%	25%	
Tax incentives	62%	33%	15%	
Technical barriers				
Security & data protection	24%	41%	35%	33% of all responses was a „high barrier"
Change control	22%	53%	25%	
Slow Internet connection	10%	66%	24%	
Local language	34%	46%	20%	

The analysis of the survey results showed that in the opinion of respondents mental barriers are more frequently than technological or legal barriers perceived as a strong obstacle in the decision on whether or not to implement cloud computing solutions. Among all responses "high barrier" as much as 38% of responses related to the mental barriers, 33% to the technical barriers and 29% to the legal barriers. Taking into account all responses, concerning twelve barriers the respondents rated the relevance of vast majority of barriers as "average". "tax incentive" was the only to be rated by most of respondents (62%) as a "low barrier".

When asked to assess the risk of cloud computing, 68% of respondents have given the answer "high risk", 23% "medium risk" and 11% "low risk".

VI. Conclusion

The results of the survey conducted among Polish SMEs confirmed the hypothesis that Polish SMEs managers perceive barriers to implementation and development of cloud computing in a similar manner to managers from other EU countries. A higher percentage of "high barrier" indications by Polish SMEs in comparison to the EU companies in the case of as many as nine barriers is a significant difference in the perception of barriers to cloud computing. Most of these differences vary from 2% (local language) to 11% (trust). In the case of "tax incentives" and "legal jurisdiction" there was 2% less indications "high barrier" in Polish SMEs than in those based in EU.

The second hypothesis i.e mental barriers, including lack of knowledge and trust are the most serious barriers to the dissemination of the cloud computing model was also positively verified. 38% of all "high barrier" indications were related to mental barriers, 33% to technical barriers and 29% to legal barriers. The biggest concerns are associated with respondents 'trust'.

Also, the third hypothesis was confirmed, as the most important legal barriers are related to concern about data security, and in particular the protection of personal data. The hypothesis that among the technical barriers "slow Internet connection" is not a significant obstacle in the development of cloud computing in the opinion of Polish managers and that this confirms a steadily increasing broad access to high-speed Internet was also verified as positive. "Slow internet conection" as "high barrier" was indicated by 24% of respondents, whereas "security & data protection" as "high barrier" was indicated by 35% of respondents. The vast majority of Polish SMEs managers evaluate the decision of implementing cloud computing as highly risky, especially in the context of losing full control over information resources and the fear of disclosing them without the owner consent.

In conclusion, the most important point is that cloud computing is still a challenge for SMEs and decisions about its usage are taken with large concerns. Managers need more knowledge about this model and about advanced tools that can facilitate their daily business activities. The vision of the benefits that cloud computing offers, and presentations of good practice are the key factors that in the near future will further increase the popularity of cloud computing in SMEs.

References

[1] S.E. Gilleta i M. Kapora *The Self-governing Internet: Coordination by Design*, MIT Press, 1996.

[2] A. Nowicki, L. Ziora, *Application of Cloud Computing Solutions in Enterprises. Review of Selected Foreign Practical Applications*, Business Informatics, No. 205, Wroclaw 2011, pp. 203-213.

[3] Q. Li, C. Wang, J. Wu, J. Li, Z.-Y. Wang, *Towards the business-information technology alignment in cloud computing environment: An approach based on collaboration points and agents*. International Journal of Computer Integrated Manufacturing, vol. 24, no. 11, pp. 1038–1057, November 2011.

[4] C. Russell, F. Jeff, J. Norm, M. Seanan, P. Carolyn, S. Patrick, J. Stanley, *Cloud Computing in the Public Sector: Public Manager "s*

Guide to Evaluating and Adopting Cloud Computing, Cisco Internet Business Solutions Group 2009.

[5] D. Jelonek, C. Stępniak, T. Turek, *The Concept of Building Regional Business Spatial Community*. In: ICETE 2013. 10th International Joint Conference on e-Business and Telecommunications. Proceedings. 29-31 July 2013, Reyklavik, Iceland 2013

[6] The National Institute of Standards and Technology's (NIST), http://www.nist.gov/itl/csd/cloud-102511.cfm (20.02.2014).

[7] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, *Above the Clouds: A Berkeley View of Cloud Computing*, Technical report No. UCB/EECS-2009-28 University of California at Berkley, 2009, USA.

[8] P. Mell, T. Grance, 2011, The NIST definition of cloud computing: Recommendations of the National Institute of Standards and Technology. Retrieved from http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[9] N. Brender, I. Markov, *Risk perception and risk management in cloud computing: Results from a case study of Swiss companies*, International Journal of Information Management 33 (2013) p. 727.

[10] V. Chang, R. J. Walters, G. Wills, *The development that leads to the Cloud Computing Business Framework*, International Journal of Information Management 33 (2013) pp. 524-538.

[11] A. Lin, N. Chen, *Cloud computing as an innovation: Percepetion, attitude, and adoption,* International Journal of Information Management 32 (2012), p. 534.

[12] P. Gupta, A. Seetharaman, J. R. Raj, *The usage and adoption of cloud computing by small and medium businesses*, International Journal of Information Management 33 (2013) pp. 861– 874.

[13] M. Greer, *Software as a service inflection point: Using cloud computing to achieve business agility*. New York: Global Authors Publishers, 2009.

[14] N. Leavitt, *Is cloud computing really ready for prime time?,* Computer, 42(1), 2009, pp. 15–20.

[15] http://www.praktycznateoria.pl/cloud-computing/

[16] *An SME perspective on cloud computing*, The European Network and Information Security Agency, 2009.

[17] *Quantitative Estimates of the Demand for Cloud Computing in Europe and the Likely Barriers to Up-take,* IDC 2012.

[18] *The activities of non-financial enterprises in 2011*, GUS, 2013.

Virtual Reality Technologies in Handicapped Persons Education

Branislav Sobota, Štefan Korečko

***Abstract*—** This paper presents a set of software tools, developed at the home institution of the authors, which utilizes virtual reality technologies in order to assist in education of handicapped persons. The tools are aimed at deaf-mute persons and can serve as sign language translators. From virtual reality-related technologies they use image recognition, data gloves and contactless sensors.

***Keywords*—**virtual reality, image recognition, education, deaf-mute persons.

I. Introduction

NOWADAYS we witness a massive penetration of information technologies into all spheres of life. This situation requires a development of more interactive and more intelligent user interfaces, which will make human - computer interaction (HCI) [1] adjusted to human users and not to the way computers operate. Such interfaces belong to the area of Human Centered Methods and Technologies (HCM-T) and Virtual Reality (VR) [2] is one of them. There is an active research in the LIRKIS laboratory at the home institutions of the authors, which focuses on the utilization of virtual reality, in the context of HCI and HCM-T, for simplification, acceleration and clarification of communication of handicapped persons with computer systems and other people. In this paper we present a set of tools, which applies VR technologies for deaf-mute persons benefit. The tools primarily deal with sign language processing.

In our work we understand the sign language as a form of communication [3]. This communication is not based on sound patterns, but on a visual transfer of meanings. The meanings can be letters, numbers, words or phrases. They are expressed by gestures, which include hand shapes, positions and movements of the hands or other body parts and facial expressions. Lexically, we can have direct gestures with clear semantics (e.g. waving, "up", "right", "I") or indirect ones where semantics is not clear at first glance (e.g. "who?"). There is no single codified version of the sign language in the world - even within one country we can have more dialects. Every country has its own vocabulary (sign register) and grammar. The tools described here have been developed for use in Slovak Republic but can be adjusted for other countries as well.

This work was supported in part by the KEGA grant no. 054TUKE-4/2013 "Application of virtual-reality technologies for handicapped persons education". In addition, the paper is a result of the Project implementation: University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS: 26220220182, supported by the Research & Development Operational Programme funded by the ERDF.

Branislav Sobota and Štefan Korečko are with the Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic (phone: +421 55 602 4313 e-mail: branislav.sobota@tuke.sk, stefan.korecko@tuke.sk).

II. Related Work and Motivation

Utilization of VR technologies in HCI context proved to be useful for handicapped persons [4]. In the case of deaf-mute persons they are mostly used to recognize the signs. However, VR is not enough for more sophisticated recognition and a translation in both ways. In such cases we need to include some artificial intelligence stuff [5]. From technological point of view the sign recognition can be implemented using contact sensors (primarily data gloves), contactless sensors or image processing technologies.

Use of data gloves represents the simplest solution. Basic gesture processing of this kind is described in [6]. Its biggest advantage is an open source platform, so it can be easily integrated to other projects. The work [7] focuses on affordability and brings a relatively cheap implementation of contact sensors. As it was mentioned above, there are many dialects of the sign language worldwide. Problems related with implementation for a specific dialect, namely a Malaysian one, are discussed in [8]. One significant disadvantage of contact sensors-based approaches is that the need to wear something (a glove) can be perceived as restrictive by humans [9], especially by the handicapped ones. This led to utilization of less limiting contact sensors based on the muscle tension sensing, such as MYO [10] or somehow simpler FIN [11]. FIN is not suitable for the sign recognition, but can be used to simplify control of electronic devices by a handicapped person.

Solutions that use contactless sensors are much less limiting and many research and development activities in this area use the Microsoft Kinect sensor [12]. A basic solution for this sensor is presented at [13] and it deals with Asian languages. The work [14] understands sign speech recognition on Kinect as recognition of individual signs. An interesting aspect of [14] is that its authors do not see a big difference between gesture recognition for deaf and mute persons and other gesture-based interfaces. Another promising solution for this kind of interface is [15], which has a potential to introduce modifications in the sign language, especially for HCI. The

ISBN: 978-1-61804-236-1

solution [16] is quite cheap and robust and focuses on a software support. The work [17] shows a potential of mobile devices in this area. Here a simple graphical application helps mute children to learn. Smartphones are also a promising platform for the both ways sign language translation.

Ideas presented in these works have been used to develop the set of tools introduced in this paper. The set can be seen as an extension of these solutions, which tries to deal with the following challenges:

- To create an implementation platform that will unite different sign language dialects. For now at least for Slovak republic.
- To deliver a software solution implemented on various platforms and using several technologies: desktop solution, mobile solution and utilization of data gloves and MS Kinect.
- To respect difficult economic situation of these persons by developing solutions as cheap as possible.
- To facilitate integration of handicapped persons into society.

III. Virtual Reality

A Virtual reality (VR) system is an interactive computer system, which creates an illusion of physical presence in a synthesized imaginary world. We can see a VR system as a tool providing a perfect simulation within the environment of tightly coupled human-computer interaction [2]. VR also includes teleoperation, telerobotics and other forms of telepresence and telecontrol.

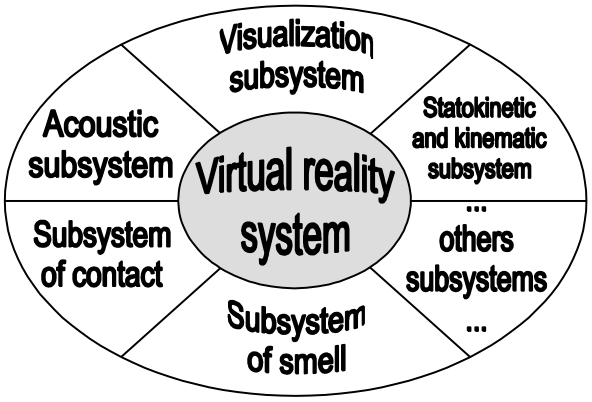


Fig. 1 Subsystems of virtual reality system

There are several levels of VR systems – from entry level, represented by ordinary personal computers, to fully immersive VR systems, which utilize special technologies [18] such as motion tracking systems, haptic devices or stereoscopic and see-through displays. A VR system can be seen as a composition of subsystems that focus on different senses or aspects of reality (Fig.1): visualization subsystem, acoustic subsystem, kinematics and statokinetic subsystem and subsystems of touch, contact and other senses (e.g. sense of smell, taste, pain or sensibility to pheromones). Which subsystems are present and how sophisticated they are depends on the level of given VR system. In addition, VR systems should be adjusted to the needs and limitations of their target groups. For example, the visualization subsystem, a standard for most VR systems, has little benefit for visually impaired persons while the subsystems of touch and contact are very important.

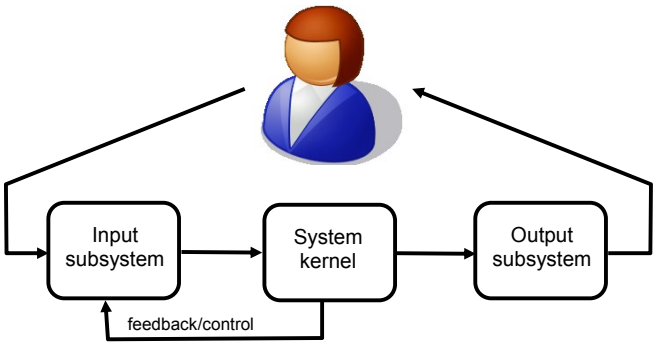


Fig. 2 Interaction in VR system

A high level VR system tries to fully immerse a user to a computer generated environment. This environment is maintained by a computer system, which graphical subsystem renders the virtual world. To achieve an effective immersion, the user's mind, and sometimes also the body, has to identify with the rendered environment. This requires a consistency between movements of the user and changes in the environment. Because the user usually only sees the virtual environment, there is no natural connection to it and such a connection have to be created. The basic system of interaction for VR is shown in Fig. 2. The quality of VR experience strongly depends on performance of underlying computer system. With increasing size and complexity of processed data and increasing output requirements more powerful and effective hardware and software is demanded. Another important trend in VR systems development is utilization of more and more interactive and intelligent user interfaces, which are more robust but also simpler from user's point of view [9]. This is also the case of handicapped persons [4], where modern information technologies can significantly help their integration to society.

IV. The Toolset

In this section we present the set of tools we developed to assist in learning and recognition of the sign language. Considering the existing approaches and solutions, a practical implementation of gesture recognition is possible by means of:

1. image recognition,
2. data gloves,
3. contactless sensors such as MS Kinect,
4. muscle tension sensing and
5. EEG scanning.

Advantages of the first method are no need of contact sensors and use of standard devices such as smartphone or

computer with a camera. Its disadvantages are a limited accuracy and a strong dependency on light conditions. The second method provides faster and more precise recognition but the need to wear additional equipment can annoy the user. The third method is fast, accurate and comfortable; however the sensors are usually not very portable and cheap. The last two approaches haven't been excessively studied and implemented at the home institution of the authors but they present promising areas for future development. For now we only experimented with single channel EEG.

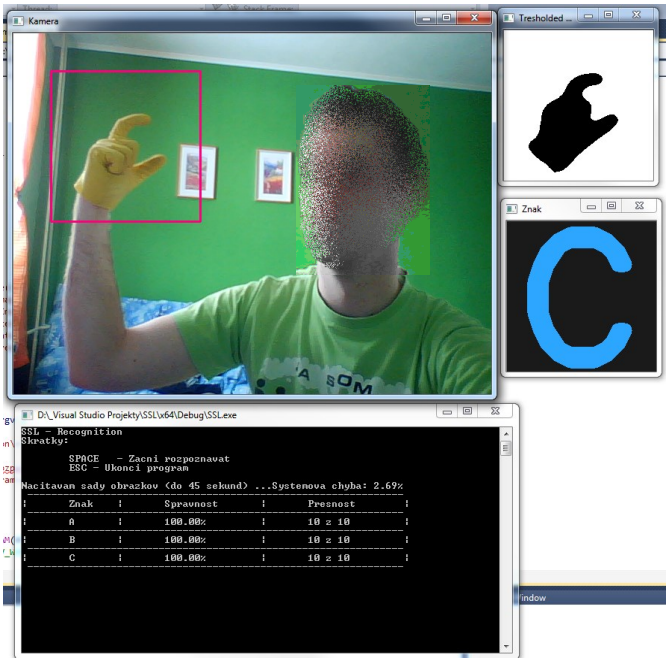


Fig. 3 Desktop image recognition-based gesture translator

A. *Image Recognition Based Gesture Translators*

The first two tools in our set use image recognition technology to identify and translate individual gestures of the sign language. First of them is a desktop application, which gesture scanning and recognition subsystem is implemented by means of the Open Source Computer Vision (OpenCV, http://opencv.org/) library. After scanning and processing a gesture it is necessary to identify it, that is to assign a meaning to it. This is done by comparison of the scanned gesture with a database of known gestures. The system performs well in recognition of static gestures (i.e. gestures that don't involve movements), but is not very useful for dynamic ones (i.e. those involving movements). Fig. 3 shows the tool in the process of scanning and recognizing a gesture for the letter "C".

The second tool uses a mobile platform, namely an Android – based mobile device. Smartphones and tablets seem to be very handy for a sign language translator because people usually carry them everywhere. Use of this prototype tool can be seen on Fig. 4. Practical experiences with the tool confirmed handiness of the platform, but also revealed recognition speed limits. These were primarily connected to limited memory and performance of the platform and a need of good network connection required for access to a shared database of gestures.

Fig. 4 Gesture recognition using mobile platform

Both tools work with bare hand but speed and accuracy of the recognition process can be increased up to about 20% when the user wears a single colored glove. This is also the case of Fig. 3 and 4.

Fig. 5 Gesture recognition with the VHand glove

B. *Data Gloves Based Tool*

The third tool allows recognizing letter gestures by means of contact sensors, namely data gloves. It supports the commercially available DGTech DG5 VHand 2.0 (Fig. 5) and our own LIRKIS glove prototype (Fig. 6). Thanks to a build-in accelerometer (both gloves) and G-sensor (LIRKIS glove) it is also possible to recognize dynamic gestures. If two gloves are available the tool is also able to recognize two-hand gestures. The recognition is faster and more precise as in the previous two tools. This makes the tool suitable for teaching of the sing

ISBN: 978-1-61804-236-1

language, including automatic testing of pupils.

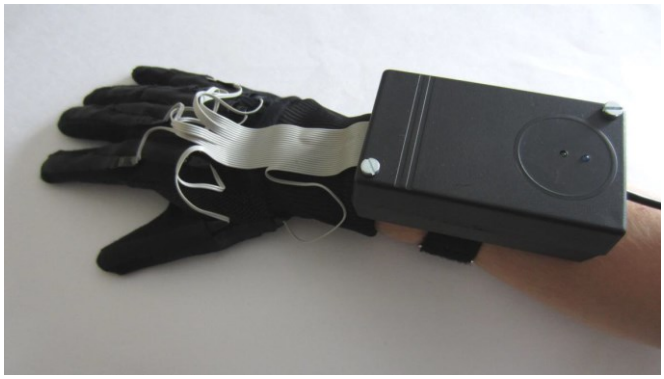
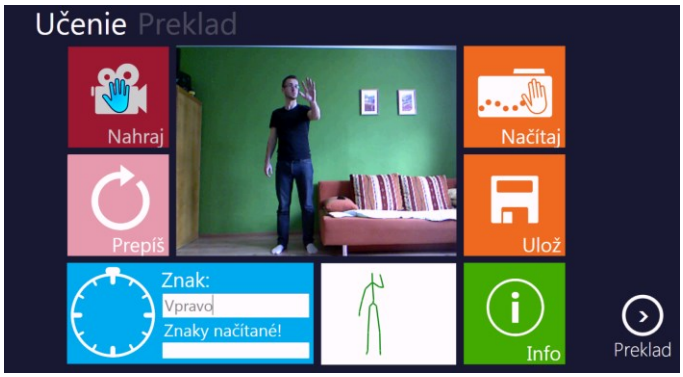


Fig. 6 LIRKIS glove prototype

C. *Contactless Gesture Recognition System*

The last tool we would like to present uses the Microsoft Kinect sensor [12] and its software development kit. We also considered the Leap Motion controller [19] but opted for Kinect because of its better availability and more widespread use in our area.



a)

b)

Fig. 7 Contactless gesture recognition tool in teaching mode: screenshot (a) and workplace (b).

The tool is a desktop application and can be run in two modes: gesture translation and gesture teaching. The second mode is used to teach the system new gestures. The Kinnect sensor not only recognizes the gestures but it can also be used for the control of the tool itself (similarly to Microsoft Xbox games). However, we are aware that even this form of control can be hard to handle for people with multiple handicaps and we plan to implement alternate means of interaction in the future.

When a user choses the teaching mode a screen with several possibilities appears (Fig. 7). Here the user can choose whether he wants to load a file with already defined gestures or to create a new file. Teaching a new gesture starts by typing its meaning in the blue box in the lower left corner of the screen. Then the user is given 3 seconds period to show the gesture. This recording period can be changed but it is sufficient for most gestures. The system records with 30 frames per second speed and each frame contains coordinates for body parts positions. These coordinates are then saved to a file and used in the translation mode for gesture recognition.

The gesture translation mode is shown in Fig. 8. Here the user shows gestures and the tool tries to recognize them on the basis of data recorded in corresponding file. As in the case of the teaching mode the most of the screen is occupied by the scanned image of the user (upper left part). The lower left part (red box) shows the meaning of a gesture just recognized, the white box next to it renders the skeleton of the user as perceived by the Kinect sensor and the right, dark grey, part hosts a list of recently recognized gestures.

Fig. 8 Contactless gesture recognition tool in translation mode

Because this version of the tool is primarily intended for pupils in the home country of the authors, its user interface (Fig. 7 and 8) is in Slovak. To better understand the interface we provide English translation in Table 1. The first column shows the Slovak word or words, used in the user interface, the second column is its English translation and the third column is more precise meaning of the word in the context of the tool.

ISBN: 978-1-61804-236-1

Table 1 Meaning of Slovak words in Fig. 7 and 8

Slovak	English	Meaning
Učenie	Teaching	Names of basic modes of the tool.
Preklad	Translation	
Nahraj	Record	Starts teaching (recording) of a new gesture.
Prepíš	Rewrite	Rewrites last recording.
Znak	Gesture	Label for a text field with gesture meaning.
Vpravo	Right	Meaning of just recorded gesture.
Znaky načítané	Gesture recorded	Message that indicates successful recording of a new gesture.
Načítaj	Load	Loading and saving of the gestures file.
Ulož	Save	
Info	Info	Shows basic info about the tool.
Výsledok	Result	Result of last gesture recognition.
Rozumieť	To understand	Meanings of recently recognized gestures.
Deň	Day	
Ahoj	Hello	

V. Conclusion

In the work presented we focused on utilization of selected virtual reality technologies and algorithms as progressive means for better, faster and easier understanding and increasing of attractiveness of handicapped persons education in the areas with hardly manageable concepts. Our particular goals have been to create an experimental setting for practical evaluation of these technologies with subsequent application of achieved results into pedagogical practice by means of supporting software and hardware solutions. The results achieved are primarily aimed at people with multiple handicaps, in particular at deaf-mute persons. Originality of the results lies in its application to the environment of Slovak republic. The tools also try to use as cheap solutions as possible in order to bring these technologies to majority of handicapped persons where the economic situation is not very good. Some of the tools developed have been field-tested at Pavol Sabadoš special boarding school in Prešov, Slovakia. In the future we plan to develop more sophisticated learning environment on the basis of these tools. To achieve this we plan to utilize artificial intelligence technologies, such as expert systems and agent systems.

References

[1] G. Sinha, R. Shahi, M. Shankar, "Human Computer Interaction," in *Proc. 3rd Int. Conf. Emerging Trends in Engineering and Technology*, IEEE, 2010, pp. 1-4.

[2] B. Sobota, Š. Korečko, F. Hrozek, "On building an object-oriented parallel virtual reality system," *Central European Journal of Computer Science*, vol. 2, no. 3, pp. 261-271, 2012.

[3] V. Vojtechovská, R. Vojtechovský, *Specific Gestures in Slovak Sign Language*. I Think – center of mute persons culture, Bratislava, 2012. 244 pp. ISBN 978-80-970601-2-1 (in Slovak)

[4] J. M. L. de Ipina et al, "Virtual reality: A Tool for the Disabled People Labour Integration," in *Proc. of Challenges for Assistive Technology*, IOS Press, 2007, pp. 141-145.

[5] B.S. Parton, "Sign language recognition and translation : A multidisciplined approach from the field of artificial intelligence," *Journal of Deaf Studies and Deaf Education*, vol. 11, no.1, pp. 94-101, 2006

[6] P.P. Abolfathi (2009), "Interpreting sign language is just the beginning for the AcceleGlove open source dataglove," *Gizmag*, [Online] available: http://www.gizmag.com/acceleglove-open-source-dataglove/12252/

[7] T. Kuroda, Y. Tabata, A. Goto, H. Ikuta, M. Murakami, "Consumer price data-glove for sign language recognition," in *Proc. 5th Intl Conf. Disability, Virtual Reality & Assoc. Tech.*, University of Reading, Oxford, UK, 2004, pp. 253-258

[8] T. T. Swee et al: "Wireless Data Gloves Malay Sign Language Recognition System", in *Proc. Of 6th Int. Conf. on Information, Communications and Signal Processing*, Singapore, Nanyang Technological University, 2007.

[9] D. Perritaz, C. Salzmann, D. Gillet, "Quality of experience for adaptation in augmented reality," in *Proc. of IEEE International Conference on Systems, Man and Cybernetic*, 2009, pp.888-893

[10] Thalmic Labs (2014), *MYO - lets you use the electrical activity in your muscles to wirelessly control your computer, phone, and other favorite digital Technologies*, [Online] https://www.thalmic.com/en/myo/

[11] RHL Vision Technologies (2014) *FIN – smart ring*, [Online] available: http://www.wearfin.com/

[12] Microsoft Corp.(2014) *MS Kinect for Windows*, [Online] available: http://www.microsoft.com/en-us/kinectforwindowsdev/Downloads.aspx; 2014

[13] Ch. Xilin et al., "Kinect Sign Language Translator expands communication possibilities", *Microsoft Research Connections*, [Online] available: http://research.microsoft.com/en-us/collaboration/focus/nui/default.aspx

[14] K.K. Biswas, S.K. Basu, "Gesture recognition using Microsoft Kinect" in *Proc. of the 5th International Conference on Automation, Robotics and Applications*, Wellington, New Zealand, 2011, pp. 100-103.

[15] S. White, D. Feng, S. Feiner, "Interaction and presentation techniques for shake menus in tangible augmented reality," in *Proc. of 8th IEEE International Symposium on Mixed and Augmented Reality*, 2009, pp.39-48.

[16] A. Agarwal, K.M. Thakur, "Sign Language Recognition using Microsoft Kinect" in *Proc. of 6th Int. Conf. on Contemporary Computing*, Noida, India, 2013, pp. 181-185

[17] D. M. Shaw, M. Patera, E. Paparidou, R. Wolff, "Evaluation of the prototype mobile phone app Pugh: a 3D cartoon character designed to help deaf children to speech read" in *Proc. 9th Intl Conf. Disability, Virtual Reality & Associated Technologies*, Laval, France, 2012, pp. 159/165

[18] F. Hrozek, "3D interfaces of systems," *Information Sciences and Technologies Bulletin of the ACM Slovakia*, vol. 5, no. 2, pp. 17-24, 2013.

[19] Leap Motion, Inc (2014): *LEAP-Motion*, [Online] available https://www.leapmotion.com/

IDEA: Security Event Taxonomy Mapping

Pavel Kácha

***Abstract*—**IDEA stands for Intrusion Detection Extensible Alert. Even though there exists a variety of models for communication between honeypots, agents, detection probes, none of them is really used because of various limitations for general usage. The paper builds upon previous work on IDEA and extends the format with taxonomies for security events and for sources and targets of attack, based on correlation of extensive body of gathered security incidents and some of existing taxonomies, and also maps unusual or too specific information into IDEA model.

***Keywords*—**alert, security event, incident response, ids, honeypot, json

I. Introduction

IDEA is an attempt to address deficiencies in automated incident report exchange. We have already defined the container for security event data in [9]. Definition of container is just one part of the job, similarly important for allowing interoperability is also definition of dictionaries for classification of various types of data, and mapping of real world data onto the IDEA format in the sane, usable way.

Creating any taxonomy, and security incident taxonomy in particular, is in no way simple task. Various users are driven by various needs and as expectations usually clash, CSIRT teams are ending up creating their own incident classifications for internal use. However, as need for more automated incident report exchange rises, and as tools for machine based security event dissemination continue to emerge, usefulness of common ground, which security teams could use at least for mapping other classifications to, becomes apparent.

Designing of security taxonomies is usually attempt to find following compromises.

A. Low level vs high level

Taxonomy may attempt to describe precise details of incident, as in venerable Howard/Longstaff1 taxonomy. The set of incident aspects and impacts is then well defined, however higher level, widely understood modus operandi (for example that incident is phishing page) is not readily obvious.

On the other hand, too vague incident types might hide important details of impact (for example – does "phishing" mean phishing spam or phishing web page? Or both?).

B. Action vs modus operandi

Incidents range from purely technical actions (connection attempt, scan) to intricate scenarios (spear phishing, social engineering), thus taxonomies have to cope with wide nature of incident complexity.

This work has been supported by the CESNET association and the operator of the Czech national research and education network referred to as CESNET2 within its "Large Infrastructure" (LM2010005) research programme, running within 2010-2015 timeframe.

Pavel Kácha works in CESNET, Zikova 4, Prague 10, Czech Republic (email: ph@cesnet.cz).

C. Exhaustive vs transparent

On the one side, incident can be classified very precisely, as for example in CAPEC [4] enumeration. However this kind of detail is usually too much of a burden to use in common scenarios. On the other side, some taxonomies use very coarse distribution, based on simplicity and ease of use (for example). For quick response security team cannot search extensive dictionary to find out meaning of very specific category. Examples of these are FICORA and CESNET taxonomies.

Incident taxonomy is usually used for classification during incident exchange and for statistical purposes. Most common statistic use case are reports and trend graphs of the most usual types of attacks, which do not need overly detailed division. Also, during incident exchange, basic incident description is usually accompanied with more detailed information if available – so there still remains possibility to use other more exhaustive specification or description of the event.

D. Rigid vs extensible

Taxonomies are usually rigid, rarely changed, causing their ageing and not being able to keep up with new types of incidents (as in Howard/Longstaff). Common ground taxonomy thus should not be static, but allow some form of extensions – be it by its authors, or by allowing side-stepping existing categories in case new incident type does not fit into predefined scenarios.

Also, sometimes one category is not enough, incidents may span more than one categories. For example security event, describing phishing email might get labelled as phishing and also as spam, because informed systems may choose to deal with incident as spam (add mail source to blacklist, learn Bayes database and so on) or specifically as phishing (add phishing URL to blacklist, inform human operator), whereas in case phishing web page gets discovered, another scenario may arise (dealing with defaced web page or poisoned DNS).

II. Existing taxonomies

There already exists a number of taxonomies, however, comparing to nowadays expectations, each of them is in various ways incomplete, outdated, or oriented to too narrow niche. Number of security teams created local taxonomies for addressing their specific needs, various security data management projects or security event detectors have their own classifications, based on their specific types of function, and of course, the real world can come up with not quite fitting security event types.

A. eCSIRT.net

eCSIRT.net taxonomy [18] is one of the most practical takes is The European CSIRT Network Incident Classification,

which is in turn based on Telia CERTCC work of Jimmi Arvidsson. Classification uses two levels, incident class and incident type. Classes are coarse grained groups, stemming from common usage, such as "Abusive Content", "Intrusions" or "Fraud", whereas types are more fine grained subclasses, such as "Spam", "Worm", or "DDoS". The structure is very practical, however taxonomy shows its age in some missing, but nowadays common security event types.

B. eCSIRT.net MkII

eCSIRT.net MkII [14] is an attempt to revive and modernize eCSIRT.net taxonomy by Don Stikvoort from SURFnet. Several missing categories are added, like non malicious events, botnet related events, and vulnerability information. We will take it as a basis for extensions and mapping.

C. Howard/Longstaff

Venerable low level take on security event classification [7] is based on splitting number of event facets ("Attacker", "Tool", "Vulnerability", "Action", "Target", "Unauthorized result", "Objectives", related to the timeline of the incident. While this attempt describes event in a great detail for recipient, it makes a great burden for sender/creator to deduce and correctly assign these facets. Also, classification shows its age and some nowadays common incident types are incorporated.

D. Longstaff at NCSC 2010

An updated model, presented at NCSC-NL International Conference 2010 [13]. Takes into consideration monetary and social information incentives of today.

E. CIF API Feed Types v1

Collective Intelligence Framework is project for gathering event data from various sources for identification, detection and mitigation (usually blacklists). It uses specific categorization for its information feeds [2].

F. CIF Taxonomy Assessment v1

Descriptive assessments of CIF [3].

G. FICORA

Categorization from Finnish Communications Regulatory Authority National Cyber Security Centre's incident submission form [12].

H. Andrew Cormack

Proposed top level classification of incidents for CSIRT teams at Terena [5].

I. SURFcert

Categorization of SURF collaborative ICT organisation for Dutch higher education and research CERT team [14].

J. CESNET-CERTS

CESNET computer security incident response team general categorization [11].

K. Warden 2

Warden is a system for sharing information about detected events, developed in CESNET. Given the types of information it supports, its classification is particularly terse [17].

L. HP TippingPoint Event Taxonomy V 2.2

Incident classification of HP flagship intrusion detection and prevention system [15].

M. CESNET Mentat

Working only with existing taxonomies would be great neglect of the real world. In the analysis and mapping we have taken into consideration types of real life events from the database of CESNET Mentat event gathering and correlation system.

N. CAPEC

Common Attack Pattern Enumeration and Classification [4] is knowledge resource database of attack mechanisms and modes operandi. It is listed here for completeness, however it was not included into mapping– it makes a great encyclopedic resource, however its vast scope and detail makes it infeasible for operative security event classification.

III. Incident Classification Mapping

For IDEA event taxonomy we have created extensive mapping between previously mentioned incident classifications. Apart from forming basis for analysis, as a side effect mapping can be readily used by security teams as a translation for communication with other parties.

Final taxonomy, based on eCSIRT.net, incident classification mapping and its discussion in this paper, comes out as follows:

Abusive
: Spam, Harassment, Child, Sexual, Violence

Malware
: Virus, Worm, Trojan, Spyware, Dialer, Rootkit

Recon
: Scanning, Sniffing, SocialEngineering, Searching

Attempt
: Exploit, Login, NewSignature

Intrusion
: AdminCompromise, UserCompromise, AppCompromise, Botnet

Availability
: DoS, DDoS, Sabotage, Outage

Information
: UnauthorizedAccess, UnauthorizedModification, UnauthorizedUsage

Fraud
: Copyright, Masquerade, Phishing, Scam

Vulnerable
: Open

Anomaly
: Traffic, Connection, Protocol, System, Application, Behaviour

Other

Test

(For full description of categories, see IDEA Classifications and Enumerations page [8].)

The whole cross reference taxonomies mapping is accessible at IDEA web page for "Incident Classification

Comparison" [10], where both eCSIRT.net taxonomies are at the very left side as the main reference, mostly because they turned out to be the most exhaustive (not counting CAPEC).

Corresponding incident type groups are clustered together where possible, and uncovered parts of taxonomies are left greyed out – or marked as catch-all category (other, unknown or similar), if particular taxonomy uses one.

If one category occupies more than one line, it means that it doesn't have counterpart in some other taxonomy.

Following chapters are discussion of this mapping and how we have come up with the result, suitable for IDEA.

IV. Discussion and implementation into IDEA

Now we are going to find parts missing or clashing among taxonomies, and try to define way to map it onto IDEA, possibly modifying MkII, to get model, which is able to convey the meaning security event for both machine and human.

Along with reasoning, examples of IDEA messages, describing related event are inserted, to verify feasibility of the result. All the messages are anonymised and stripped to bare minimum, but still perfectly valid.

A. Blacklists, whitelists

Information about being put into blacklist/whitelist is quite commonly communicated information – one is not able to process all and every blacklist/whitelist on the wild, moreover various lists and databases pop up and disappear frequently. People often rely on getting this information from third party sources, aggregators, etc.

Whitelists are either lists of addresses, knowingly clean in some particular aspect, or attempts to monetize on impression of legitimacy of certain company's email or internet assets (DNSWL), or site/vendor/organization specific exception lists, not relevant to security event dissemination.

Blacklists specifically important to security teams are those, which inform about vulnerabilities and specific security problems – lists of WWW pages, injected with phishing or malware, open relay mailservers, open recursive resolvers, etc.

In incident handling process, these are usually communicated in the same way as locally found vulnerabilities, with additional specifics accompanying the message.

These events can be in MkII represented as basic "Vulnerability", and if used at incident message, by additional labelling specific to transport protocol and/or format and/or concerned parties needs – I believe narrow categories akin to Phishing WWW, Malware WWW, Open Relay Mailserver are out of scope of such a general categorization.

Found in:

- CIF API Feed Types v1: infrastructure/whitelist, domain/whitelist, email/whitelist, url/whitelist
- CIF Taxonomy Assessment v 1: Whitelist
- HP Tipping Point: IP Filters/Deny, IP Filters/Accept

IDEA representation:

```
{
  "Format": "IDEA0",
  "ID": "c34bf422-931c-4535-9c6b-257128185265",
  "DetectTime": "2014-11-03T10:33:12Z",
  "Category": ["Vulnerable.Open"],
  "Confidence": 0.5,
  "Description": "Open Recursive Resolver",
  "Source": [
    {
      "Type": ["Open"],
      "IP4": ["93.184.216.119"],
      "Proto": ["udp", "domain"]
    }
  ]
}
```

B. Anomalies

Anomalies, such as excessive traffic, might later be identified as security problem (for example DoS or DDoS), however they might end up as accidental peak or outage, or completely innocent. As anomalies can be important to security teams as indicator of possible attack, or as a correlation element in investigation, I think these should be taken into account in security events transfer. I see two possibilities to represent them:

- specific top level category, for example *Anomaly*, with suitable subcategories, I'd suggest *Traffic*, *Connection*, *Protocol*, *System*, *Application*, *Behaviour*
- when anomaly arises, we usually have suspicion, which types of incidents can it cause (excess traffic → DOS, overlaid TCP packets → exploit, too many connections → dictionary attack, etc.). So there is possibility to use these deduced categories, but for incident handling we might allow another dimension – certainty of detection (or self trust). However, that requires support from underlying transport format.

Both of these approaches have its use, first is usable, when we are not able to connect possible situation with any type of attack, whereas second describes situation, which could potentially evolve into real threat, or get recognized as such by closer analysis.

Found in:

- HP Tipping Point: Traffic Thresholds, Application or Protocol Anomaly

IDEA representation in Anomaly category:

```
{
  "Format": "IDEA0",
  "ID": "4390fc3f-c753-4a3e-bc83-1b44f24baf75",
  "DetectTime": "2014-02-01T18:32:03Z",
  "Category": ["Anomaly.Traffic"],
  "ConnCount": 3352,
  "Description": "Possible DoS",
  "Source": [
    {
      "IP4": ["93.184.216.119"],
      "Proto": ["tcp"]
    }
  ]
}
```

IDEA representation as suspicion:

```
{
  "Format": "IDEA0",
```

```
  "ID": "0bd857b6-7c4d-4a17-ad1a-bcb1cc8eaa6b",
  "DetectTime": "2014-02-01T19:28:23Z",
  "Category": ["Availability.DoS"],
  "ConnCount": 3352,
  "Confidence": 0.5,
  "Description": "Possible DoS",
  "Source": [
    {
      "IP4": ["93.184.216.119"],
      "Proto": ["tcp"]
    }
  ]
}
```


C. Backscatter/Bounce

Bounce is distinct flavour of spam – DSN messages generated by servers in reaction to non deliverable spam messages with forged sender, thus sent to innocent forged recipients. That might validate another category. However mechanism of backscatter – forging sender data – is more general and abused also in DDOS attacks, like DNS amplification or various other types of UDP reflection attacks, which might indicate that this information should be represented or communicated differently/orthogonally, possibly as facet of the source.

Found in:

- CESNET CERTS: Bounce

IDEA representation:


```
{
  "Format": "IDEA0",
  "ID": "bf8344d7-a0da-4724-92da-ccda382d7e72",
  "DetectTime": "2014-01-03T01:23:42Z",
  "Category": ["Abusive.Spam"],
  "Description": "Spam bounce",
  "Source": [
    {
      "Type": ["Spam", "Backscatter"],
      "IP4": ["93.184.216.119"],
      "Proto": ["tcp", "smtp"]
    }
  ]
}
```


D. Scans

Number of existing taxonomies distinguish between specific types of IP based reconnaissance, the basic observed types being host scan, port scan, service scan, application scan, port sweep, ICMP probe. This again denotes technical facet of the attack, which can be communicated by some other means – in security event description formats for example by type of network and application protocol used, and number of ports and machines scanned.

Some taxonomies also differentiate events based just on cardinality of attack – singular events might get marked akin to "connection attempt". In fact there is no way to be sure, whether singular events are part of greater reconnaissance or not, without additional information usually from other sources. Most important information, which this distinction conveys, is the severity of the attack, and that's also orthogonal information, which should get communicated by other ways.

Found in:

- HP Tipping Point: Reconnaissance or Suspicious Access
- Warden 2: Portscan, Probe
- Mentat: Probe, Portscan, Connection attempt, Ping probe, SYN/ACK scan or DOS attack

IDEA representation:


```
{
  "Format": "IDEA0",
  "ID": "3ad275e3-559a-45c0-8299-6807148ce157",
  "DetectTime": "2014-03-22T10:12:56Z",
  "Category": ["Recon.Scanning"],
  "ConnCount": 633,
  "Description": "Ping scan",
  "Source": [
    {
      "IP4": ["93.184.216.119"],
      "Proto": ["icmp"]
    }
  ]
}
```


E. Vulnerabilities

Various event detectors are also able to deduce attacked application or even name of the exploit used. That however also does not belong into general taxonomy, as this usually goes along as additional info – and there is number of well known databases of vulnerabilities, which can be used.

Found in:

- Mentat: EPMAPPER exploitation attempt, SMB exploitation attempt, SQL query attempt, URL attack attempt, Webattack, Open recursive resolver
- HP Tipping Point: Vulnerability

IDEA representation:


```
{
  "Format": "IDEA0",
  "ID": "3ad275e3-559a-45c0-8299-6807148ce157",
  "DetectTime": "2014-03-22T10:12:31Z",
  "Category": ["Recon.Scanning"],
  "ConnCount": 633,
  "Description": "EPMAPPER exploitation attempt",
  "Ref": ["cve:CVE-2003-0605"],
  "Source": [
    {
      "IP4": ["93.184.216.119"],
      "Proto": ["tcp", "epmap"],
      "Port": [24508]
    }
  ],
  "Target": [
    {
      "Proto": ["tcp", "epmap"],
      "Port": [135]
    }
  ]
}
```


F. Botnets

Botnets are one of the most common threats today. Taxonomies sometimes differentiate at least between C&C servers and worker drones, because bringing down C&C servers is of more benefit, than cleaning up workstation infected by drone. Importance of this information might validate adding new category, however it's again more of a

technical facet. When integrating taxonomy into security event format, this information should not be omitted, at least as severity of the incident, or as a property of attack source, also with indication of fastflux possibility.

Found in:

- CIF API Feed Types v1: infrastructure/botnet, url/botnet, domain/botnet, infrastructure/fastflux, domain/fastflux
- CIF Taxonomy Assessment v1: Botnet, Fastflux
- Mentat: Botnet Drone, Botnet Proxy, Botnet_c_c

IDEA representation:

```
{
  "Format": "IDEA0",
  "ID": "cca3325c-a989-4f8c-998f-5b0e971f6ef0",
  "DetectTime": "2014-03-05T15:52:22Z",
  "Category": ["Intrusion.Botnet"],
  "Description": "Botnet Command and Control",
  "Source": [
    {
      "Type": ["Botnet", "CC"],
      "IP4": ["93.184.216.119"],
      "Proto": ["tcp", "ircu"],
      "Port": [6667]
    }
  ]
}
```

G. Phishing/Pharming/Scam

At least one examined taxonomy distinguishes between phishing and pharming – that's also technicality, which should be identifiable from accompanying information (cache poisoning, DNS break-in, etc.).

However, well known type of incidents are variation on Nigerian 419 scam. That might fit into "Abusive Content/Spam" category, but that does not tell the whole story – it's not *just* spam. It might also fit into "Fraud/Masquerade" category, but that depends on what designers of eCSIRT.net taxonomy exactly mean by "masquerade" – whether posturing as specific person (identity theft), or general con (variation of social engineering). I suggest adding "Fraud/Scam" category for clarity.

Found in:

- CESNET CERTS: Phishing, Pharming, Scam

IDEA representation:

```
{
  "Format": "IDEA0",
  "ID": "9729ea4a-a260-40c0-8e63-0cb0b2687177",
  "DetectTime": "2014-02-22T13:35:03Z",
  "Category": ["Fraud.Scam"],
  "Description": "419 mail scam",
  "Source": [
    {
      "Type": ["Spam"],
      "IP4": ["93.184.216.119"],
      "Proto": ["tcp", "smtp"]
    }
  ]
}
```

H. Suspicious

URLs found in spam messages or in sandboxed malware binaries may or may not be necessarily evil. They are definitely suspicious, but spammers and malware creators often incorporate innocent URLs to lure automated tools astray. I am not convinced of the necessity of new specific category, in security event messages this information will go under "Abusive Content/Spam" or "Malicious Code", and extracted URL should be marked as unclear by other means (specific type, reliability).

Found in:

- Mentat: Sandbox URL, Spam URL

IDEA representation:

```
{
  "Format": "IDEA0",
  "ID": "4d52640a-5363-497a-a7d9-bcbde759cb7d",
  "DetectTime": "2014-02-21T16:01:32Z",
  "Category": ["Abusive.Spam"],
  "Description": "Spam URL reference",
  "Source": [
    {
      "Type": ["OriginSpam"],
      "URL": ["http://www.example.com/"],
      "Proto": ["tcp", "http", "www"]
    }
  ]
}
```

I. Searches

During reconnaissance, attackers often use Google searches ("Google Hacking"), or conduct various suspicious searches against company sites. This activity can be detected, either by Google aimed project (Google Hack Honeypot [6]) or by local IDS systems. This type of information gathering does not precisely fit into any MkII subcategory, I suggest adding "Information Gathering/Searching" category.

Found in:

- CIF Taxonomy Assessment v1: Searches

IDEA representation:

```
{
  "Format": "IDEA0",
  "ID": "b7dd112c-9326-49e6-a743-b1dce8b69650",
  "DetectTime": "2014-02-13T02:21:15Z",
  "Category": ["Recon.Searching"],
  "Description": "Suspicious search",
  "Source": [
    {
      "IP4": ["93.184.216.119"],
      "Proto": ["tcp", "http", "www"]
    }
  ],
  "Target": [
    {
      "URL": ["http://www.example.com/search=%20union%20select
%20password%20from%20users%20%2D%2D"]
    }
  ]
}
```

J. Local

At least one taxonomy incorporates breaches into company policies. As these can be local specific, they don't belong into general taxonomy. In IDEA, these can be represented by locally defined nodes, as IDEA container is freely extensible.

Found in:

- HP Tipping Point: Security Policy

K. Unclassifiable

The situations may arise, where we are aware of wrongdoing, but are not able to classify it by means of existing taxonomy class. There are two possible scenarios:

1. We don't know what exact type of incident that is, and what particular class it belongs to, maybe because we need additional information to find out. We can then use educated guess (and possibly, if channel allows for that, add certainty of that guess), or it might again warrant "Anomaly" category.
2. We know the type of incident and it's completely new one, which does not fit into any of the existing categories. We can either use Other, or at least top level category (if it does fit into one). Or we can aim for extensibility and leave creating of new subcategories on users – and codify them later into standard based on what is experienced in the wild.

V. IDEA Implementation

A. Security event taxonomy

eCSIRT.net MkII comes out as the most comprehensive, yet still practical solution. From mapping and comparison with other taxonomies and several real world incidents we have implemented following updates:

1. Adding "Anomaly" category, with following subcategories (incident examples): Traffic, Connection, Protocol, System, Application, Behaviour (see IV.K).
2. Add "Scam" incident example into "Fraud" (see IV.G).
3. Add "Searching" incident example into "Information Gathering" (see IV.I).
4. Don't stay rigid, allow side-stepping, make taxonomy extensible by users (see I.D).
5. Allow multicategorization, where applicable (see I.D).

B. Source/target specifics taxonomy

Along with security event taxonomy, we have added attack source/target classification, which stems partially from the need to complement MkII with more specific information about attack origin or destination. Some of the source/destination types were already used in examples, however full list is (along with whole adapted and abbreviated MkII) at [8]. These classification names are meant to be used as (possible multiple) tags in "*Source.Type*" or "*Target.Type*" field of IDEA messages.

VI. Conclusion

IDEA is an attempt to address deficiencies in automated incident report exchange. In this report we have created mapping of various incident taxonomies to each other, identified some practical deficiencies and omissions in most recent of them – MkII, and recommended and implemented modifications. We have also created auxiliary classification of attack sources/destination and shown real world examples to verify feasibility. The whole specification is available at [8].

Taxonomy mapping is also readily usable for translation between classification in various security team, thus simplifying teams work [10].

With this updates IDEA is able to reasonably encompass majority of information from other taxonomies, and describe all security events we have encountered so far.

References

[1] *Warden* [online]. CESNET. Copyright 2010-1013. Last updated 17 April 2013. Available: http://warden.cesnet.cz

[2] *CIF API Feed Types v1* [online]. Cited 13 February 2014. Available: https://code.google.com/p/collective-intelligence-framework/wiki/API_FeedTypes_v1

[3] *CIF Taxonomy Assesment v1* [online]. Cited 13 February 2014. Available: https://code.google.com/p/collective-intelligence-framework/wiki/TaxonomyAssessment_v1

[4] *Common Attack Pattern Enumeration and Classification* [online]. MITRE. Cited 13 February 2014. Available: http://capec.mitre.org

[5] A. Cormack, *Proposed top level classification of incidents* [online]. TERENA. Cited 13 February 2014. Available: http://www.terena.org/activities/tf-csirt/pre-meeting3/TLversion0_2.html

[6] *Google Hack Honeypot* [online]. Honeynet Alliance. Cited 13 February 2014. Available: http://ghh.sourceforge.net/

[7] J. D. Howard, T. A. Longstaff, *A Common Language for Computer Security Incidents*. Sandia National Laboratories, October 1998. SAND98-8667. Available: http://www.cert.org/research/taxonomy_988667.pdf

[8] P. Kácha, *IDEA/Classifications and Enumerations* [online], CESNET. Cited 13 February 2014. Available: https://csirt.cesnet.cz/IDEA/Classifications

[9] P. Kácha, IDEA: "Designing the Data Model for Security Event Exchange", *17th International Conference on Computers: Recent Advances in Computer Science*, Rhodos, 16 July 2013, ISBN: 978-960-474-311-7, ISSN: 1790-5109.

[10] P. Kácha, *Incident Classification Comparison (with eCSIRT.net mkII as main reference)* [online], CESNET, 10 January 2014. Available: https://csirt.cesnet.cz/IDEA/Classifications?action=AttachFile&do=get&target=Incident+classification+comparison.ods

[11] P. Kácha, *OTRS: CSIRT WorkFlow Improvements* [online]. CESNET. August 2010. Available: http://archiv.cesnet.cz/doc/techzpravy/2010/otrs-csirt-workflow/

[12] E. Koivunen, *Effective Information Sharing for Incident Response Coordination*. Aalto University, 30 May 2010. Available: http://personal.inet.fi/koti/erka/Studies/DI/DI_Erka_Koivunen.pdf

[13] T. Longstaff, *Where the Wild Things Are* [online]. NCSC-NL International Conference 2011. Available: https://www.ncsc.nl/binaries/en/conference/conference-2011/speakers/tom-longstaff/1/TomLongstaffPresentation.pdf

[14] D. Stikvoort, *Incident Classification* [online]. 23 May 2013. Available: http://www.terena.org/activities/tf-csirt/meeting39/20130523-DV1.pdf

[15] *TippingPoint Event Taxonomy, Version 2.2* [online]. Cited 13 February 2014. Available: http://hitec.com.do/carlos/CarlosMeza/__Curso_IPS/ProductDocumentation/TECHD94-EventTaxonomy.pdf

[16] P. Vachek, "CESNET Audit System", *Proceedings of the 13th WSEAS International Conference on COMPUTERS*, Rodos Island, July 23-25, 2009, ISBN: 978-960-474-099-4.

[17] *Warden archive* [online]. CESNET. Cited 13. March 2014. Available: ftp://homeproj.cesnet.cz/tar/warden/warden-client-2.1.tar.gz

[18] *WP4 Clearinghouse Policy* [online]. eCSIRT.net. © 2002-2003 by PRESECURE Consulting GmbH, Germany. Available: http://www.ecsirt.net/cec/service/documents/wp4-clearinghouse-policy-v12.html#HEAD6

A parallel algorithm for optimal job shop scheduling of semi-constrained details processing on multiple machines

Daniela I. Borissova and Ivan C. Mustakerov

***Abstract*—**The paper presents an approach for a variant of constrained job shop scheduling where processing of some details is independent and other have fixed processing order (semi-constrained scheduling). The described approach aims to determine a schedule that minimizes the total makespan in such way that all given operations sequences are satisfied. For the goal, a parallel algorithm is proposed based on linear programming optimization tasks that are solved in parallel. The described approach for optimal job shop scheduling of semi-constrained details is numerically tested for real job shop scheduling problem.

***Keywords*—**Job shop scheduling, linear programming, minimal makespan, semi-constrained details processing.

I. Introduction

The scheduling is a key factor for manufacturing productivity. Effective manufacture scheduling can improve on-time delivery, reduce inventory, cut lead times, and improve the utilization of bottleneck resources [1].

One of the most studied combinatorial optimization problems is the job shop scheduling problem. Nevertheless, it still remains a very challenging problem to solve optimally. From a complexity point of view, the problem is NP-hard i.e. it can be solved in nondeterministic polynomial time [2], [3].

The simplest scheduling problem is the single machine sequencing problem [4]. Minimizing the total makespan is one of the basic objectives studied in the scheduling literature. The shortest processing time dispatching rule will give an optimal schedule in the single machine case if the tool life is considered infinitely long [5]. The scheduling with sequence-dependent setups is recognized as being difficult and most existing results in the literature focus on either a single machine or several identical machines [6]-[8]. The real-life scheduling problems usually have to consider multiple no identical machines.

D. I. Borissova is with the Institute of Information and Communication Technologies at Bulgarian Academy of Sciences, Sofia – 1113, Bulgaria, Department of Information Processes and Decision Support Systems (phone: 3952 9792055; e-mail: dborissova@iit.bas.bg).

I. C. Mustakerov is with the Institute of Information and Communication Technology at the Bulgarian Academy of Sciences, Sofia – 1113, Bulgaria, Department of Information Processes and Decision Support Systems (phone: 3952 9793241; e-mail: mustakerov@iit.bas.bg)).

Most of the processing machines needed to process the jobs are available in the manufacturer's own factory and are of fixed (finite) number. Sometimes, certain details must be ordered to a third party companies to complete very specific processing as molding for example. In cases like that, the processing schedules are to be agreed for delivery times from the third-party processing. That means generating a schedule to process all jobs, so as to minimize the total cost, including the satisfaction of the due dates of the jobs [9]. Different manufacturing environments induce different scheduling constraints, some of which may be very specific to the problem under consideration [10].

The classical job shop scheduling problem is one of the most typical and complicated problems formulated as follows: 1) a job shop consists of a set of different machines that perform operations of jobs; 2) each job is composed of a set of operations and the operation order on machines is prescribed; 3) each operation is characterized by the required machine and the processing time. In the last two decades, numerous techniques was developed on deterministic classical job shop scheduling, such as analytical techniques, rule-based approach and meta-heuristic algorithms and algorithms using dynamic programming [11]-[15].

Approximately up to 2004 the computers have had gradually increasing of CPU performance by increasing of operating frequency, and the need of multi core systems was not so obvious. NVIDIA has invented the graphics processing unit (GPU) that became a pervasive parallel processor to date. It has evolved into a processor with unprecedented floating-point performance and programmability and today's GPUs greatly outpace CPUs in performance, making them the ideal processor to accelerate a variety of data parallel applications. GPUs have hundreds of processing cores and with CUDA programming model [16] software and hardware architecture is available using of a variety of high level programming languages. This represented a new way to use the GPU as a general purpose parallel computer processor. This opens up new horizons in development and application of new approaches based on parallel algorithms [17].

The proposed scheduling approach concerns a problem of scheduling for multiple details with fixed processing time and predetermined order of processing operations over different machines. An essential feature of the investigated job shop

scheduling problem is that: 1) the processing of some details depends on processing of other details i.e. a group of details have predetermined order of processing and 2) the processing of the other parts is independent of each other. This variant of job shop scheduling can be named as semi-constrained job shop scheduling problem. It is approached in the paper by means of an algorithm based on parallel solving of a number of integer linear programming tasks. The main goal is to determine a schedule that minimizes the total makespan in such way that all details processing conforms to the given restrictions. The proposed parallel algorithm for optimal job shop scheduling of semi-constrained details processing on multiple machines is numerically tested for a real life example.

II. Problem Description

There is a group of details that need to be processed on multiple machines. Some of these details are connected with each other through given order of processing while other can be processed in any order. All details have predetermined sequence of operations on different machines. The details processing times on machines are deterministic and are known in advance. The problem is to determine the minimum makespan for all details processing according to requirements.

For clarity of presentation the investigated job shop problem will be explained by a real life example for a set of six details (jobs) with given sequences of operations that should be processed on four different machines with known processing time on each machine. All available data are summarized in Table I where operations' designation O_{ij} means processing of detail i on machine j and processing times are given in hours.

TABLE I
Input Data for Details Processing

Details (Jobs)	Operations	Processing time on M1	Processing time on M2	Processing time on M3	Processing time on M4
D_1	O_{11}	8			
	O_{12}		6		
	O_{14}				6
D_2	O_{21}	8			
	O_{22}		9		
	O_{24}				6
D_3	O_{31}	8			
	O_{33}			8	
	O_{32}		8		
D_4	O_{41}	4			
	O_{42}		2		
	O_{43}			2	
D_5	O_{51}	4			
	O_{52}		9		
	O_{53}			5	
D_6	O_{61}	6			
	O_{63}			4	

The sequence of operations for each detail are given as D_1 $\{O_{11}, O_{12}, O_{14}\}$, D_2 $\{O_{21}, O_{22}, O_{24}\}$, D_3 $\{O_{31}, O_{33}, O_{32}\}$, D_4 $\{O_{41}, O_{42}, O_{43}\}$, D_5 $\{O_{11}, O_{12}, O_{13}\}$ and D_6 $\{O_{61}, O_{63}\}$. Due their post-processing specifics the details D_4, D_5 and D_6 should be processed in a sequential order. All jobs cannot overlap on the machines and one job cannot be processed simultaneously by two or more machines. Each operation needs to be processed during an uninterrupted period of a given length on a given machine. The goal of the investigated scheduling problem is to determine a schedule that minimizes the total makespan. The described problem can be represented as machine-oriented Gantt chart visualizing the sequence of details processing as shown in Fig. 1

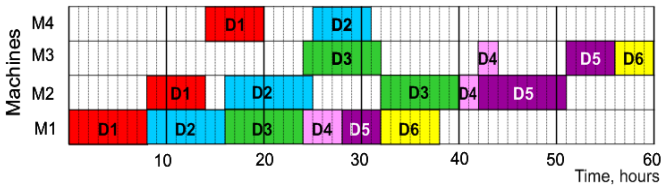


Fig. 1. Gantt chart for a schedule

III. Mathematical Model Formulation

Most variants of job shop scheduling problem are NP-hard in the strong sense and thus defy ordinary solution methods. That is why new techniques are required to overcome difficulties and to be applied to particular manufacturing job shop scheduling problems. The generalized goal of most of optimal scheduling problems is to minimize the overall costs. Although many costs could be considered for optimization, the minimizing of details processing time duration is one of most frequently used. It provides the effective machines utilization and serves the optimization of details delivering and storage. The overall details processing time duration (makespan) can be defined as difference between end processing moment of the last detail and start processing moment of the first detail and if the processing starts at moment zero moment then the objective can be minimization of the end processing moment of the last detail. Using those considerations, an optimization model can be formulated following the notations:

1) number of details indexed by $i \in \{1,2,\ldots,N\}$
2) number of machines for detail processing, indexed by $j \in \{1,2,\ldots,M\}$
3) job processing times $T_{i,j}$ of each detail i on machine j are known constants.
4) $x_{i,j}$, is the moment of time for starting of processing of detail i on machine j.

The scheduling problem is formalized via linear programming formulation that minimizes the makespan as:

$$\min \rightarrow \sum_{i=1}^{N} x_{i,end} \tag{1}$$

subject to

$$x_{i,j+1} - x_{i,j} \geq T_{i,j}, \ \forall i=1,2,\ldots,N, \tag{2}$$

$$x_{i+1,j} - x_{i,j} \geq T_{i,j}, \ \forall j=1,2,\ldots,M \tag{3}$$

$$x_{i,j} \geq 0 \tag{4}$$

ISBN: 978-1-61804-236-1

The objective function (1) minimizes the processing end time of all details. The relation (2) expresses the restriction for operation sequence for each detail, while relation (3) illustrates the restriction for the details processing order (if any). For dependant details processing there exist a certain order of processing, but in case of independent details the processing order is not fixed. This way formulated model can be used to determine the optimal makespan for a given sequence of details processing. To find the optimal makespan among all of the possible sequences of independent details processing a parallel algorithm can be applied.

IV. Parallel Algorithm for Minimal Makespan Determination

To find the optimal scheduling that is minimal in the sense of shortest overall makespan, a parallel algorithm for optimal job shop scheduling of semi-constrained details processing on multiple machines is developed as shown in Fig. 2.

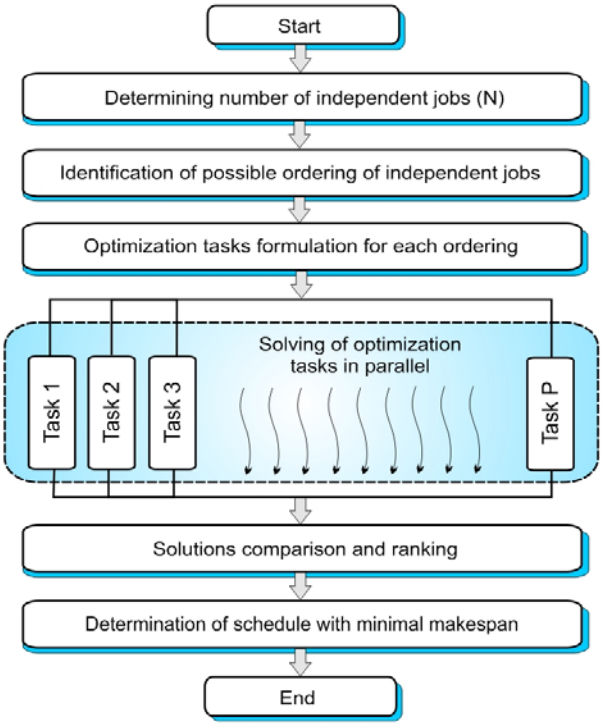


Fig. 1. Parallel algorithm for job shop scheduling

On the first step of the algorithm the independent jobs are to be defined and designated using name of the corresponding detail. If exist dependant jobs (as for details D4, D5 and D6) they are considered as one independent job named after first detail of processing sequence (D4). Then the overall number of independent details is determined. On the second step all possible orderings (permutations) for processing of independent details are defined. On the next step each of the details processing ordering is formalized by proper optimization task following the model (1) – (4). It is important to stress here that all formulated in this way optimization tasks are independent of each other. The solution of any of them does not depend on data or solution of other tasks. This makes them perfect candidates for using of parallel threads for their solving on step 4. Then, tasks solution results (makespan values) are compared and ranked. On the last step the schedule corresponding to the task with best solution with minimal makespan value is chosen as optimal job shop schedule.

V. Numerical Example

In deterministic job shop scheduling problem, is assumed that all processing times are fixed and known in advance, so using the input data from Table I, the optimization model (1) – (4) can be expressed as:

$$\min\left(x_{1,end} + x_{2,end} + x_{3,end} + x_{4,end} + x_{5,end} + x_{6,end}\right) \quad (5)$$

$$x_{1,2} - x_{1,1} \geq 8 \quad (6)$$

$$x_{1,4} - x_{1,2} \geq 6 \quad (7)$$

$$x_{1,end} - x_{1,4} \geq 6 \quad (8)$$

$$x_{2,2} - x_{2,1} \geq 8 \quad (9)$$

$$x_{2,4} - x_{2,2} \geq 9 \quad (10)$$

$$x_{2,end} - x_{2,4} \geq 6 \quad (11)$$

$$x_{3,3} - x_{3,1} \geq 8 \quad (12)$$

$$x_{3,2} - x_{3,3} \geq 8 \quad (13)$$

$$x_{3,end} - x_{3,2} \geq 8 \quad (14)$$

$$x_{4,2} - x_{4,1} \geq 4 \quad (15)$$

$$x_{4,3} - x_{4,2} \geq 2 \quad (16)$$

$$x_{4,end} - x_{4,3} \geq 2 \quad (17)$$

$$x_{5,2} - x_{5,1} \geq 4 \quad (18)$$

$$x_{5,3} - x_{5,2} \geq 9 \quad (19)$$

$$x_{5,5} - x_{5,3} \geq 4 \quad (20)$$

$$x_{5,end} - x_{5,5} \geq 5 \quad (21)$$

$$x_{6,3} - x_{6,1} \geq 6 \quad (22)$$

$$x_{6,end} - x_{6,3} \geq 4 \quad (23)$$

- to restrictions for the details priority processing:

$$x_{2,1} - x_{1,1} \geq 8 \quad (24)$$

$$x_{3,1} - x_{2,1} \geq 8 \quad (25)$$

$$x_{4,1} - x_{3,1} \geq 8 \quad (26)$$

$$x_{5,1} - x_{4,1} \geq 4 \quad (27)$$

$$x_{6,1} - x_{5,1} \geq 4 \quad (28)$$

$$x_{2,2} - x_{1,2} \geq 6 \quad (29)$$

$$x_{3,2} - x_{2,2} \geq 9 \quad (30)$$

$$x_{4,2} - x_{3,1} \geq 8 \quad (31)$$

$$x_{5,2} - x_{4,2} \geq 2 \quad (32)$$

$$x_{4,3} - x_{3,3} \geq 8 \quad (33)$$

ISBN: 978-1-61804-236-1

$$x_{5,3} - x_{4,3} \geq 2 \quad (34)$$

$$x_{6,3} - x_{5,3} \geq 5 \quad (35)$$

$$x_{2,4} - x_{1,4} \geq 6 \quad (36)$$

$$x_{i,j} \geq 0 \quad (37)$$

The formulated task (5) – (37) takes into account the details processing sequence $D_1 \rightarrow D_2 \rightarrow D_3 \rightarrow D_4 \rightarrow D_5 \rightarrow D_6$. To define the minimum makespan for other processing sequence this task should be reformulated. The group of restrictions for details priority processing (24) to (36) has to be changed to correspond to other possible details processing sequence. There are 3 details (D1, D2 and D3) that can be processed in any order. The group of dependant details D4, D5 and D6 can be considered as one independent detail and the number of all possible processing sequences can be calculated as number of permutations of 4, i.e. number of different processing sequences that have to be evaluated is equal to 4! = 24.

For example, if details processing sequence is $D_1 \rightarrow D_3 \rightarrow D_2 \rightarrow D_4 \rightarrow D_5 \rightarrow D_6$ the restrictions (24) – (26) should be reformulated as:

$$x_{3,1} - x_{1,1} \geq 8 \quad (24)$$

$$x_{2,1} - x_{3,1} \geq 8 \quad (25)$$

$$x_{4,1} - x_{2,1} \geq 8 \quad (26)$$

The objective function (5) and the rest of restrictions remain the same.

If details processing sequence is $D_2 \rightarrow D_1 \rightarrow D_3 \rightarrow D_4 \rightarrow D_5 \rightarrow D_6$ the restrictions (24) – (26) have to be changed as:

$$x_{1,1} - x_{2,1} \geq 8 \quad (24)$$

$$x_{3,1} - x_{1,1} \geq 8 \quad (25)$$

$$x_{4,1} - x_{3,1} \geq 8 \quad (26)$$

and again the objective function and the rest of restrictions remain the same.

If the group of details D_4, D_5 and D_6 is to be processed in the first place i.e. details processing order is $D_4 \rightarrow D_5 \rightarrow D_6 \rightarrow D_1 \rightarrow D_2 \rightarrow D_3$ the restrictions (24) – (36) are transformed to:

$$x_{5,1} - x_{4,1} \geq 4 \quad (24)$$

$$x_{6,1} - x_{5,1} \geq 4 \quad (25)$$

$$x_{1,1} - x_{6,1} \geq 6 \quad (26)$$

$$x_{2,1} - x_{1,1} \geq 8 \quad (27)$$

$$x_{3,1} - x_{2,1} \geq 8 \quad (28)$$

$$x_{5,2} - x_{4,2} \geq 2 \quad (29)$$

$$x_{1,2} - x_{5,2} \geq 9 \quad (30)$$

$$x_{22} - x_{12} \geq 6 \quad (31)$$

$$x_{3,2} - x_{2,2} \geq 9 \quad (32)$$

$$x_{5,3} - x_{4,3} \geq 2 \quad (33)$$

$$x_{6,3} - x_{5,3} \geq 5 \quad (34)$$

$$x_{3,3} - x_{6,3} \geq 4 \quad (35)$$

$$x_{2,4} - x_{1,4} \geq 6 \quad (36)$$

with the same objective function and the remaining restrictions.

In similar way, all possible combinations of detail processing sequences can be reflected in 24 different modifications of basic optimization task (5) – (37). The solving of all of the tasks can be done in parallel because all of the tasks are entirely independent of each other. The result of the solutions is 24 job shop schedules corresponding to different details processing sequences.

VI. Results Analysis and Discussion

The solutions of all optimization tasks corresponding to all possible combinations for details processing sequences along with their total makespan are shown in Table II.

The makespan values in tasks solutions vary within interval of 65 to 52 hours for different details processing sequences. Among them the optimal one with minimal makespan equal to 52 hours is for details processing sequence: $D2 \rightarrow D1 \rightarrow D3 \rightarrow D4 \rightarrow D5 \rightarrow D6$.

The corresponding schedules for each processing sequence are illustrated in Fig. 3. For the described example, 11 different makespans have been distinguished that could not be determined by intuitive considerations. Increasing the number of independent details will increase the number of processing sequences and therefore the number of tasks that must be solved but because of parallel algorithm for solution of each task, this will not affect the computational complexity. Despite the fact that integer problems are difficult to solve (in general they are NP-hard), the formulated optimization problems and numerical testing show quite acceptable solution times of few seconds by means of LINGO solver [18].

All real-life job shop scheduling problems have their own specifics. When analyzing the resulting schedules it can be seen, the relationship between the processing details sequence and machines occupation have a significant impact on overall manufacturing process performance. For the described example, it turned out that Machine 1 is the busiest machine among the others. One possible approach to shorten the overall makespan is to consider more than one machine of type 1 and to estimate the influence of machine's number on the total makespan.

The proposed approach based on parallel solution of a set optimization tasks can be used for other similar problems concerning optimal job shop scheduling.

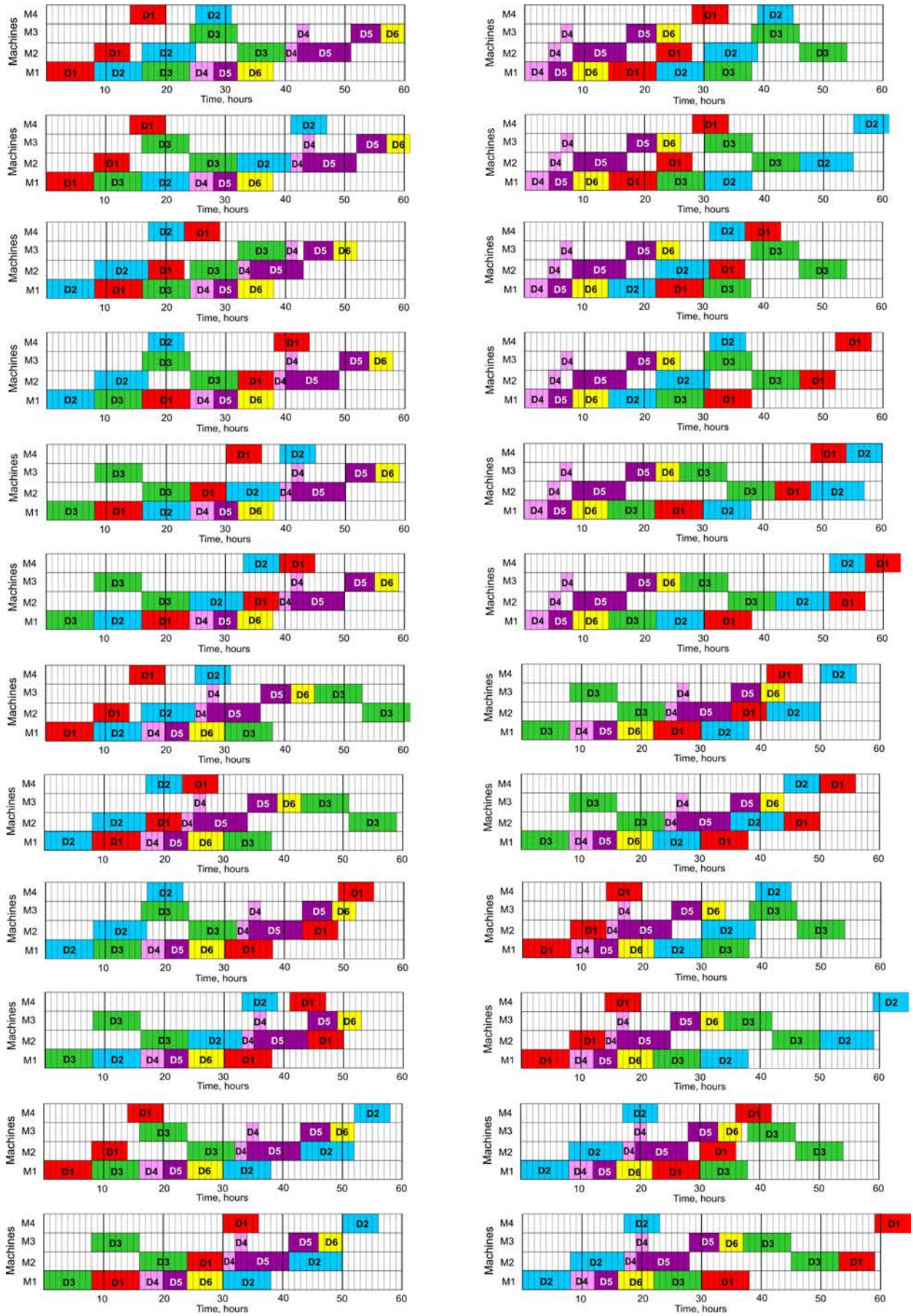


Fig. 3. Schedules for different details processing sequences

TABLE II
SOLUTIONS RESULTS

Sequence of the details processing	Total makespan, hours
D_1, D2, D3, D4, D5, D6	60
D1, D3, D2, D4, D5, D6	61
D2, D1, D3, D4, D5, D6	**52**
D2, D3, D1, D4, D5, D6	58
D3, D1, D2, D4, D5, D6	59
D3, D2, D1, D4, D5, D6	59
D4, D5, D6, D_1, D2, D3	54
D4, D5, D6, D1, D3, D2	61
D4, D5, D6, D2, D1, D3	54
D4, D5, D6, D2, D3, D1	58
D4, D5, D6, D3, D1, D2	60
D4, D5, D6, D3, D2, D1	63
D1, D2, D4, D5, D6, D3	61
D2, D1, D4, D5, D6, D3	59
D3, D2, D4, D5, D6, D1	55
D2, D3, D4, D5, D6, D1	53
D1, D3, D4, D5, D6, D2	58
D3, D1, D4, D5, D6, D2	56
D3, D4, D5, D6, D1, D2	56
D3, D4, D5, D6, D2, D1	56
D1, D4, D5, D6, D3, D2	54
D1, D4, D5, D6, D2, D3	65
D2, D4, D5, D6, D1, D3	54
D2, D4, D5, D6, D3, D1	65
Minimal makespan: 52 hours	

VII. CONCLUSION

In this paper, a deterministic job shop scheduling approach for details processing on multiple machines based on integer linear programming model is described. The goal of described job shop scheduling is to determine the minimum makespan for a number of semi-dependant details (some with independent processing and other with dependant of each other processing) with different operations on different machines. To find the minimum of total makespan, a number of identical optimization tasks corresponding to all permutations of independent details processing sequences are formulated. The main contribution of the paper is using of the developed model in an algorithm based on solving of all formulated tasks in parallel. The execution of the algorithm provides a set of job shop optimal schedules for all possible details processing sequences. Then the best schedule and corresponding processing sequence in sense of minimal makespan are determined.

As extensions and future investigations, a possible direction is to explore how increasing number of identical machines will influence the algorithmic and computational difficulties.

For large scale job shop problems, where the total makespan could be essentially bigger, this approach can contribute not only to reduce the makespan via schedules optimization, but also to decrease the overall production time and costs.

ACKNOWLEDGMENT

The research work reported in the paper is partly supported by the project AComIn "*Advanced Computing for Innovation*", grant 316087, funded by the FP7 Capacity Programme (Research Potential of Convergence Regions).

REFERENCES

[1] D. Chen, P. B. Luh, L. S. Thakur and J. Moreno Jr., "Optimization-based manufacturing scheduling with multiple resources, setup requirements, and transfer lots", *IIE Transactions*, vol. 35, 2003, pp. 973-985.

[2] M. R. Garey, D. S. Johnson, and R. Sethi, "The complexity of flowshop and job shop scheduling", *Mathematics of Operations Research*, vol. 1, no. 2, 1976, pp. 117-129.

[3] Yu. N. Sotskov and N. V. Shakhlevich. "NP-hardness of shop-scheduling problems with three jobs", *Discrete Applied Mathematics*, vol. 59, 1995, pp. 237-266.

[4] S. J. Mason, P. Qu, E. Kutanoglu and J. W. Fowler, "The single machine multiple orders per job scheduling problem", available: http://ie.fulton.asu.edu/files/shared/workingpapers/MOJ_Paper.pdf

[5] M. S. Akturk, J. B. Ghosh and E. D. Gunes, "Scheduling With Tool Changes to Minimize Total Completion Time: A Study of Heuristics and Their Performance", *Naval Research Logistics*, vol. 50, no. 1, 2003, pp. 15-30.

[6] S. C. Kim and P. M. Bobrowski. "Impact of sequence-dependent setup time on job shop scheduling performance". *Int. Journal of Production Research*, vol. 32, no. 7, 1994, pp. 1503-1520.

[7] I. M. Ovacik, and R. Uzsoy, "Rolling horizon algorithm for a singlemachine dynamic scheduling problem with sequence-dependent setup times", *Int. Journal of Production Research*, vol. 32, no. 6, 1994, pp. 1243-1263.

[8] H. L. Young, K. Bhaskaran, and M. Pinedo, "A heuristic to minimize the total weighted tardiness with sequence-dependent setups", *IIE Transactions*, vol. 29, no. 1, 1997, pp. 45–52.

[9] J. Wang, P. B. Luh, X. Zhao and J. Wang, "An Optimization-Based Algorithm for Job Shop Scheduling", *Sadhana*, vol. 22, 1997, pp. 241-256.

[10] P. Baptiste and C. L. Pape, "Disjunctive constraints for manufacturing scheduling: principles and extensions", *Int. Journal of Computer Integrated Manufacturing*, vol. 9, no. 4, 1996, pp. 306-310.

[11] M. Pinedo, "Stochastic scheduling with release dates and due dates", *Operations Research*, vol. 31, no. 3, 1983, pp. 559-572.

[12] R. R. Weber, P. Varaiya and J. Walrand, "Scheduling jobs with stochastically ordered processing times on parallel machines to minimize expected flowtime", *Journal of Applied Probability*, vol. 23, no. 3, 1986, pp. 841-847.

[13] D. Golenko-Ginzburg and A. Gonik, "Optimal job-shop scheduling with random operations and cost objectives", *Int. Journal of Production Economics*, vol. 76, no. 2, 2002, pp. 147-157.

[14] S. R. Lawrence and E.C. Sewell, "Heuristic optimal static and dynamic schedules when processing times are uncertain", *Journal of Operations Management,* vol. 15, no. 1, 1997, pp. 71-82.

[15] J. A. S. Gromicho, J. J. van Hoorn, F. Saldanha-da-Gama and G. T. Timmer. "Solving the job-shop scheduling problem optimally by dynamic programming", *Computers & Operations Research,* vol. 39, 2012, pp. 2968-2977.

[16] NVIDIA. Whitepaper NVIDIA's Next Generation CUDA Compute Architecture: Fermi, 2009.

[17] S. A. Mirsoleimani, A. Karami and F. Khunjush. "A Parallel Memetic Algorithm on GPU to Solve the Task Scheduling Problem in Heterogeneous Environments", *Genetic and Evolutionary Computation Conference*, 2013, pp. 1181-1188.

[18] Lindo Systems ver. 12, http://www.lindo.com

Mining Precise Typestates by Exploring Benefits of Available Specifications

Yi Zhang(Naval Academy of Armament, Beijing, China, 100036)
Ge Chang(Naval Academy of Armament, Beijing, China, 100036)
Yazhuo Dong(Naval Academy of Armament, Beijing, China, 100036)
yi_nirvana@163.com

***Abstract*—**Typestate is a formalism to specify correct sequences of method invocations. Despite their usefulness, typestates are often unavailable in practice because writing them is cumbersome and error-prone. The state abstraction approach is proposed by researchers to automatically mine typestates. The main idea is to use abstract values of object fields to label states during the mining process. However, current approaches simply abstract values of object types to *null* or *not-null*, which results in overly general typestates that include many erroneous behaviors. In this paper, we propose to mine precise typestates of the composite object by leveraging states of its field objects. These states are encoded within available specifications of field objects. Instead of the *not-null* state for a field object, we distinguish different states of a field object specified in its specifications and use them to construct abstract states of the composite object. Our approach makes the states of field objects visible to the miner and improves the precision of mined typestates by effectively increasing the number of their states. The empirical evaluation shows that our approach can significantly improve the precision of mined typestates by removing erroneous behaviors and without noticeable loss of recall.

***Keywords*—**object-oriented typestate, FSM model, the state abstraction approach. the empirical evaluation.

I. INTRODUCTION

IN object-oriented programs, programmers write code by invoking various application programming interfaces (APIs). In general, not all method invocation sequences are legal. There are constraints on the temporal order of method invocations. For example, programmers should not *write* into a file after it has been *closed*. API protocols specify which API method call sequences are allowed, which are very useful in many software engineering activities. They can aid the generation of test cases [1]. Program analysis tools can use them to find certain errors [2][3][4]. In addition, formal specifications including API protocols can support the understanding of correct software behaviors, which is central to software maintenance.

As writing API protocols is cumbersome and requires expert knowledge of corresponding APIs, they are often missing, incomplete or out-of-date despite their usefulness. Mainstream object-oriented languages provide only informal documentation to support API protocols. To address this problem, researchers have developed specification mining techniques to mine API protocols from API client programs [5][6][4][7][8]. These techniques typically produce API protocols in the form of finite state machines (FSM). Unfortunately, FSMs produced by these mining techniques are not suciently precise for practical use. Mined FSMs are often overly general and include many spurious behaviors. The imprecision becomes especially worse when mined FSMs are large and complex [9]. Even for small, two-state FSMs, the false positive (false specifications) can be high (e.g., 90-99% [4], and 63% when precision and recall is balanced [10]). Overgeneralized FSMs can hinder the effectiveness of downstream analysis, verification and validation techniques by producing many false negatives and/or positives. To tackle this problem, Researchers try to validate mined specifications [11].

Typestate [12] is a formalism intended to capture API protocols. The observation behind typestates is that whether a method invocation is available on an object depends on not only the type of the object but also its internal states. Techniques to mine typestates based on explicit object states [7][13][14][15] have been proposed. The main idea of this approach is to use values of object fields to label states during the mining process. Compared with other automatabased FSM mining approaches [5][16][17] that resort to heuristics to identify equivalent states, this approach naturally labels equivalent states with the same field values. This leads to a straightforward mining approach with much less time complexity that is typically linear to the size of the input trace (In contrast, the commonly used *kTail* algorithm [18] and typical PFSA learners [5] have the running time quadratic and cubic to the length of the input trace, respectively). This approach is also expected to have less overgeneralization. However, to avoid produce too large and under-generalized models, abstract instead of concrete field values are used to label states. Unfortunately, choosing an appropriate abstraction function at the right abstraction level for typestates mining is diffcult and the current adopted abstraction function *abs* is as follows: values of reference fields (objects and arrays) are abstracted to *null* (=*null*) or *not-null* (≠*null*), values of numerical fields are abstracted to *larger than zero* (>0), *less than zero* (<0), or *equal to zero* (=0), and values of boolean fields remain unchanged. However, this state abstraction approach can produce typestates including so much

This work is supported by NSFC of China under the NO. 60903057.

overgeneralization that they are too imprecise to be used as specifications [19].

In this paper, we mine typestates for objects in the form of FSMs, one FSM model for each field of the object with public methods of the object accessing (read/write) this field as the alphabets. Such models consider unrelated methods in isolation and can lead to more complete models for the object [6]. We observe that one main source of the overgeneralization of result FSMs is the *null-abstraction* of reference fields that views possibly many different states of a field object as a single state ≠null. This abstraction for reference fields are too coarse. It actually assumes that API specifications of all field objects are the simple one-state automaton with the form of the Figure 1 (a). Such an automaton specifies that invocations of instance methods should be made on an initialized object, and nothing else. Obviously, many important properties of the field object are missed from Figure 1 and typestates mined under this abstraction have a small number of states and much nondeterminism. For example, Figure 1 (b) presents the model of the **BufferedOutputStream** mined by the null-abstraction approach. This FSM is useless and it includes the spurious behavior that the stream can be written into after it has been closed. Figure 1 (c) presents the API specification of the **OutputStream**. The close of the **BufferedOutputStream** calls the close of the **OutputStream** which transits the state of the out from open to closed. However, this state change is invisible to the miner because these two states are abstracted into the same state ≠*null*.

We also observe that many techniques mine typestates from scratch, ignoring some potential existing helpful information. In object-oriented programming, the *composition* is one of the most common ways to construct new classes from existing libraries. Because of their long-term usage and well-understanding, existing libraries possibly have some formal specifications that are either specified by their developers /users or mined by some specification miners, considering the fact that specification mining techniques have made important progress after its more that a decade's development [20][21]. Existing well-known specifications include the resource-releasing specifications like the one presented in Figure 1 (c), the *collection-iterator* specification and so on. The simple null-abstraction of field objects for typestates mining has two main drawbacks: first, the mined typestates are overly general; second, the mined typestates of the composite object can contradict specifications of its field objects, resulting in self-incompatible specifications of the entire program.

In this paper, we propose to leverage available specifications of field objects to mine precise typestates of composite objects through the state abstraction approach. Instead of the single ≠ *null* state of the null-abstraction, we distinguish different states of a field object encoded in its specifications and use them to construct abstract states of the composite object. When available specifications of the field object are finite state properties, we monitor the field object against these properties and use states that are reached as abstract states of the field object. These abstract states are used to label states of the composite object during the mining process. For example, Figure 1 (d) presents the enhanced model of the **Buffered-OutputStream** by leveraging the specification of the **Output-Stream** presented in Figure 1 (c). The **close** invocation of the **BufferedOutputStream** calls the **close** of the **OutputStream** which transits the state of the field *out* from *open* to *close*. In this way, the states *open* and *closed* of the field *out* make the *open* state and the *closed* state in Figure 1 (d) distinguishable and the overgeneralization is removed.

Dallmeier et al. [15] propose to automatically generate test cases to enrich mined typestates. This approach is effective to increase the number of transitions but has a limited power of discovering new states due to the state abstraction function used. For example, the generated test cases may cover the transition of theclose from the state *open* to the state *closed* in Figure 1 (d), but the mined model is still the one presented in Figure 1 (b) because these two states are indistinguishable. Their approach and ours complement each other. The approach to mine deep models is presented by Dallmeier et al. [22]. For the state abstraction, they do not simply map fields of reference types to =*null* or ≠*null*, but consider the fields of the composite object' reference fields. To mine the typestates in Figure 1 (d), this approach depends on the existence of a state-indicating field in the field *out* or in its descendant fields, which can not be guaranteed in general. Generally speaking, we do not know what is the best abstraction function for specific fields without further knowledge. So, simply to iteratively abstract the fields of a reference field are not appropriate. In contrast, if there are already available specifications for a field object, we should utilize it.

We empirically validate our approach through comparing typestates mined by the null-abstraction and typestates mined by our approach. We mined typestates for classes in several packages of the Java system library and use traces generated from the DaCapo benchmark programs [23]. The evaluation shows that our approach can significantly improve the precision by removing erroneous behaviors from mined typestates. Meanwhile, no recall is lost for the cases of our benchmark programs. Our miner is fast and the overhead introduced by monitoring field objects is limited: the time increases are between 3-16%.

The rest of this paper is organized as follows. Section 2 gives the background of object-oriented typestates and explanations of some terms used in the paper. Section 3 discusses our approach in detail. Section 4 presents the experimental evaluation of our approach. Section 5 discusses related work and section 6 concludes.

II. Terminology and Background

In this section, we first give the background of object-oriented typestates and then define some terms used in the paper.

A. Object-Oriented Typestates

Object-oriented typestate systems are proposed in the literature [24][25]. Because typestates reflect how state changes of objects can a effect valid method invocations, a typestate is an abstraction over concrete object states and can be characterized

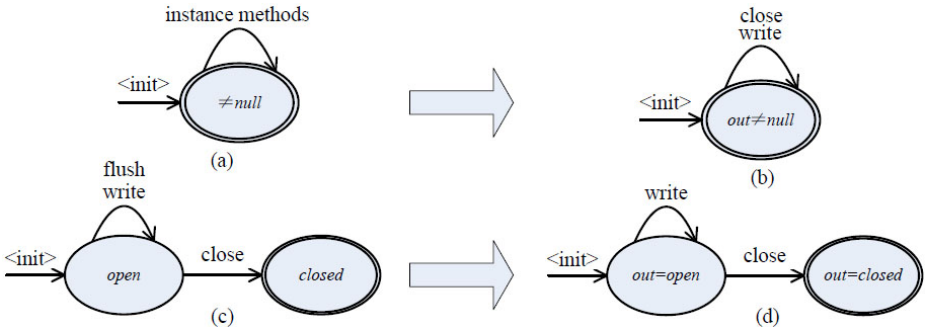


Fig. 1 API specification of an object under the null-abstraction (a) and mined model for the **BufferedOutputStream** by state-of-the-art state abstraction approach (b); Available specification of the **OutputStream** (c) and mined model for the **BufferedOutputStream** by our approach (d).

$$
\begin{aligned}
M &:= C \mid M \wedge M \\
C &:= V \rightarrow V \\
V &:= (s_1, \ldots, s_n)
\end{aligned}
$$

Fig. 2 A simplified method specification language for the typestates of the object-oriented programs.

by the values of all fields of an object. Typestates are mapped onto the fields of the class by defining a predicate for each typestate, called a *state invariant,* which can be any boolean combination of state tests, state comparisons, integer comparisons, boolean constants and fields. The substitutability of subtypes for super types is preserved by the state refinement that a subtype can define a set of sub-states as special cases of an existing state. The specification of a method can be changed through the method *refinement*. A method can be re-specified more precisely in a subtype based on the refined sub-states.

The main role of typestates is to specify methods. Figure 2 gives a simplified method specification language. A method is specified with an intersection of cases, which means that all these cases hold. A *case* represents a state transition which is denoted as $A \rightarrow B$ to express that a method requires a source state A and produces a destination state *B*. The source state is a vector consisting of the states of the receiver object and its arguments (in their order in the signature). The destination state has one more state for the method's return if any. Nondeterminism of state transitions can be expressed using the intersections of different cases. For example, $A \rightarrow B \wedge A \rightarrow C$ represents that starting at the state *A*, executions of a method can transit to the state *B* or the state *C*. The state invariant is evaluated to test whether an object is in a particular state. Either statically checked [24] or dynamically checked [25], state invariants of typestates are evaluated for every method invocation: source state violations are flagged as precondition violations and destination state violations are flagged as postcondition violations. Source states and destination states are actually treated as preconditions and postconditions of corresponding methods, respectively.

B. Definitions

We first define the *trace* of program executions that we use to mine typestates and then give the definition of the *typestate model*, or simply *model*, that our approach produces as the API protocol of a class.

Definition 1. (**Trace**) We distinguish two types of objects in the input traces. The *composite objects* are objects whose specifications we intend to mine, and the *field objects* are objects that are assigned to fields of composite objects during runtime. A *trace* $T = \langle e_1, \ldots, e_n \rangle$ is a sequence of events, where an event *e* can be a field assignment event *f a*, or a method entry event *en*, or a method exit event *ex*. A field assignment event is a tuple $f\ a = \langle cn, f, cs \rangle$, where a value *cn* is assigned to the field *f* of a composite object *cs*. *cn* is a subcomponent object if *f* is of reference type. For primitive types (numerical types and boolean types), their concrete values are recorded in field assignment events. A method entry event is a tuple $en = \langle m, o, F \rangle$ with *F* as a set of fields of *o*, where the method *m* of the object *o* begins to execute and *m* accesses fields in *F* during this invocation. A method exit event is a tuple $ex = \langle m, o \rangle$, where the method *m* of the object *o* completes its execution and exits. If an event *e* appears in the trace *T*, we write $e \in T$.

Because we are only interested in objects, calls to staticmethods are excluded from the trace. So, the receiver object of a method entry/exit event always exists. We mine specifications of a composite object from the viewpoint of the object's user (caller of the object's public methods), and thus require that method entry/exit events do not appear in the trace if the method is called by a method with the same receiver object. For multi-threaded applications, we create a separate trace file for each thread.

Figure 3 presents a trace fragment. Each line corresponds to an event and contains the identifier of the event at the beginning of the line. The event en_6 indicates the beginning of the execution of the **write** of the **BufferedOutputStream** object 657, and the event ex_6 indicates the end of this execution. During this execution, the **write** first calls the **write** of the **FileOutputStream** object 647 (en_7 and ex_7), then assigns the value 0 to the field count of the object 657 ($f\ a_6$), then calls the **flush** of the object 647 (en_8 and ex_8), and at last calls the **close** of the object 647 (en_9 and ex_9).

Our aim is to mine the API protocol of a class. Compared with the object-oriented typestate systems introduced above,

ISBN: 978-1-61804-236-1

our typestate model has several adaptations. First, we omit states of the parameters of a method invocation and only consider the state of its receiver object. This omission can significantly reduce the complexity of the mining process and will produce a conservative result. Second, the mined model is the set of all valid method invocation sequences and the states in the model is anonymous, that is, we omit state invariants or state labels. Such a model can provide a succinct representation of API protocols. Third, we create a model for each field of the a class in the intuition that methods are unrelated unless they access the same variable [6]. Each model consists of public methods that access a particular field of the class. The separation of methods includes only related methods into a model and can produce more complete models for the class.

en_1 FileOutputSteam.<init>, FileOutputSteam:647
ex_1 FileOutputSteam.<init>, FileOutputSteam:647
en_2 BufferedOutputSteam.<init>, BufferedOutputSteam:657, {out, buf}
fa_1 FileOutputSteam:647, out, BufferedOutputSteam:657
fa_2 byte[]:658, buf, BufferedOutputSteam:657
ex_2 BufferedOutputSteam.<init>, BufferedOutputSteam:657
en_3 BufferedOutputSteam.write, BufferedOutputSteam:657, {count, buf}
fa_3 int:1, count, BufferedOutputSteam:657
ex_3 BufferedOutputSteam.write, BufferedOutputSteam:657
en_4 BufferedOutputSteam.write, BufferedOutputSteam:657, {count, buf, out}
en_5 FileOutputSteam.write, FileOutputSteam:647
ex_5 FileOutputSteam.write, FileOutputSteam:647
fa_4 int:0, count, BufferedOutputSteam:657
fa_5 int:1, count, BufferedOutputSteam:657
ex_4 BufferedOutputSteam.write, BufferedOutputSteam:657
en_6 BufferedOutputSteam.close, BufferedOutputSteam:657, { count, buf, out}
en_7 FileOutputSteam.write, FileOutputSteam:647
ex_7 FileOutputSteam.write, FileOutputSteam:647
fa_6 int:0, count, BufferedOutputSteam:657
en_8 FileOutputSteam.flush, FileOutputSteam:647
ex_8 FileOutputSteam.flush, FileOutputSteam:647
en_9 FileOutputSteam.close, FileOutputSteam:647
ex_9 FileOutputSteam.close, FileOutputSteam:647
ex_6 BufferedOutputSteam.write, BufferedOutputSteam:657

Fig. 3 Fragment of the trace used to mine the model in Figure 1 (d).

Definition 2. (**Typestate Models**) A *typestate model M* for a class *c* is a collection of sub-models M_f, each of which is a FSM with its alphabet as the set of all public methods of *c* that access the field *f* of *c*. A behavior *T* (sequence of method invocations) of an object of *c* is valid if $T \downarrow M_f$ is accepted by every sub-model M_f of *c*. $T \downarrow M_f$ is a sub-sequence of T that keeps in their original order only method invocations whose method appears in the alphabet of M_f.

III. Approach

In this section, we present our approach in detail. The input includes the execution traces of training programs and available specifications of field objects. Any finite-state properties in the form of FSMs can be fed into our approach. Monitoring field objects and mining typestates of composite objects proceed concurrently. The field objects in the traces are monitored against their specifications. The states that are reached are fed into the typestate miner as the abstract states for the corresponding field objects. The output is the mined typestate models of composite objects.

A. Monitoring Field Objects

Given the finite-state specification of a field object, we should monitor the sequence of method invocations on the field object in the trace to determine the abstract states of the field object between method calls in the trace. The specification of a field object considered in this paper can be any FSM with its alphabet as the set of public methods of the field object. We create a monitor for each field object with input specifications. The monitor simulates the trace on the specification automaton, advancing the automaton by one step when a public method invocation on the object is encountered in the trace. For example, if we monitor the **FileOutputStream** object 647 in the trace in Figure 3 against the specification in Figure 1 (c), we will get the sequence of state transitions of the object 647 as follows: (en_1, ex_1) open (en_5, ex_5) open (en_7, ex_7) open (en_8, ex_8) open (en_9, ex_9) closed.

As FSM are commonly nondeterministic automata, more than one states may be simultaneously reached during monitoring. If the set of reached states are used as the abstract state of the field object, we need some criterion to identify same states of the composite object. We can naively devise a few criteria, one of which is that if two sets of states of the field object have at least one same state, the two states corresponding to these two set are equivalent and the merged result abstract state is the union of these two sets. Because it is difficult to choose among various such criteria, we adopt a different solution that is more straightforward. We require that input FSMs are deterministic since deterministic and nondeterministic automata are equivalent. If input FSMs are nondeterministic, we first transform them to equivalent deterministic FSMs. In this way, There is a single state that is reached at any time during monitoring. This single state is used to label states of the composite object and two states of the field object are equivalent if and only if they are the same one.

Although we expect that input specifications of field objects are reliable, the requirement of full complete and precise specifications is not practical and will limit the applicability of our approach. In addition, the trace can include erroneous behavior in cases that there are some events in the trace that violate input specifications but do not cause the execution of the program to fail. These cases can manifest themselves as violations of input specifications during monitoring. When the monitor of a field object encounters a violation, we set the current state of the field object as $\neq null$ from now on and stop the monitoring of the field object while the mining proceeds. In addition, the monitored object, the violated specification and the violating trace is logged. Such logs can be used to detect potential bugs in the program and/or enhance input specifications.

B. Mining Typestate Models

The intuition of typestate miners through state abstraction is straightforward: for each method invocation of the target object in the trace, we add a transition with labelled source and destination states to the model of the object. The first and last method invocations of the object in the trace identify the initial and final states, respectively. We use the same state abstraction

function abs as that of state-of-the-art typestate miners [22][13] to compute abstract states of fields. However, instead of the single $\neq null$ state, the abstract state of a field object is the state of its monitor that is currently reached. If a field has the value *null* or there are not specifications for its value, the null-abstraction is used.

The methods of a field object may be called by a method whose receiver is not the composite object of this field object. For example, a field object can be returned by a getter method and then is accessed by a method of another object. In these cases, two consecutive method calls of the composite object in the trace may return (the first method call) and enter (the second method call) in different states of the field object. This causes

Algorithm: MiningTypestates

Input:
$TS = \{t_1, \ldots, t_n\}$: Types of composite objects, optional
T: Trace
$SP = \{S_1, \ldots, S_k\}$: Deterministic FSMs for field objects
Output:
$MS = \{M_t \mid t \in TS\}$: Typestates for composite objects
L: logs of input specification violations

```
1:  foreach e ∈ T in the order in T do
2:      if e = <m, o, F> is method entry ∧ t(o) ∈ TS then
3:          foreach f ∈ F do
4:              add a transition tran(e,f) for m from abs(o.f) to null to M(t,f)
5:      if e = <m, o> is method exit ∧ t(o) ∈ TS then
6:          foreach f ∈ entry(e).F do
7:              set destination state of tran(entry(e),f) as abs(o.f)
8:      if e = <m, o, F> is method entry ∧ S(t(o)) ∈ SP then
9:          if a monitor mo exists for o then
11:             advance mo by one step
12:             if a violation encountered then
13:                 set abstract state of o to ≠null
14:                 destroy mo
15:                 log this violation
16:         update composite object' state with current state of mo
17:     if e = <v, f, o> then
18:         if f is of reference type and S_f ∈ SP then
19:             if v has a monitor mo then
20:                 update state of o with current state of mo
21:             else if v is newly created then
22:                 create a monitor mo for v
23:                 update state of o with current state of mo
24:             else update state of o with abs(v)
25:         else update state of o with abs(v)
26:     for all t ∈ TS do
27:         MS_t = union of all MS_o with o of type t for each field
28:     return MS
```

problems for the labeling of conceptually equivalent states of the composite object. The naive solution to this problem is as follows. we define the label for an abstract state of a field object as a pair $<l_1, l_2>$, where l_1 is the current abstract state of the field object after a method call of the composite object and l_2 is the current abstract state of the field object before the next method call of the composite object in the trace. Two such labels are equivalent if and only if the first states and the second states are same, respectively. If the specification of a field object has n different states, there can be n^2 different labels. On one hand, under such a labeling scheme, the miner can produce complex typestates with too many states. On the other hand, we observe that accesses to a field object from other objects than its composite object are out of control of the composite object and they can vary greatly in different contexts. So, we adopt a different solution. When the miner encounters a method call on the field object made by a method whose receiver is not its composite object, the state of this field object is abstracted to $\neq null$ and monitoring of this field object terminates. This scheme can provide a balance between the complexity and expressiveness of mined typestates.

In a field assignment event $<v, f, o>$ with f of reference type, v may be in any of its states. To determine the abstract states of v before and after method calls of its composite object o, we must begin to monitor v once after it is created. The naive approach is to monitor these objects that are bound to get assigned to a field of composite objects. However, to decide the the future field assignment of a potential field object when it is created, we must traverse the trace once before the mining and tag these objects. Because the traces are usually very long, traversing traces is very time-consuming. Here, we adopt a comprised approach that we monitor all potential field objects that have input specifications once after they are created. However, because only some of them will be assigned to fields of composite objects in common cases, this approach will cause unnecessary overhead. We give another fast approach as follows. We tag every object when it is created as a newly created object. If a method entry event with a *newly created* object as the receiver is encountered then after in the trace, the tag of this object is removed. During mining, when a field assignment event with a not null field object is encountered, we create a monitor for the field object if this field object has not a monitor but has the *newly created* tag, or abstract it to $\neq null$ and do not monitor it if it has not a monitor and has not the tag. This approach can monitor most field objects that are accessed only by their composite objects. The fast approach may omit to monitor some field objects but avoid the unnecessary overhead to monitor non-field objects. These two approaches do not need a pre-traversal of the traces.

The algorithm **MiningTypestates** shows the pseudo code to mine precise typestates of composite objects. The input includes *TS* as types of target objects for which we intend to mine typestates, *T* as a trace of the execution of the training program, and *SP* as a set of specifications of the deterministic FSM form for field objects. The input of *TS* is optional. If it is not specified, the miner infers typestates of all objects except for those with inputs specifications in *SP* in the trace. For space limit, the initialization of data structures is omitted. We define the function *entry*: {*method exit events*} → {*method entry events*} that maps a method exit event to is corresponding method entry event, and the function *t*: {*objects*} → {*types*} that maps an object to its runtime type. We write *e.l* to denote the element *l* (such as *F* for a method entry event *e*) of the event *e*.

This algorithm processes events in the trace one by one. For each method entry event of the target object, we add a transition to the sub-model for each field access by the method invocation (lines from 2 to 4). The destination state of a transition is determined at the corresponding method exit event (lines from 5 to 7). For a method entry event of the field object with input

specifications, we advance the monitor of the field object if any by the called method and log violations if any (lines from 8 to 16). For a field assignment event, we create a monitor for the field object if it has input specifications but has not a monitor now (lines from 17 to 25). To choose which objects to monitor, we incorporate the fast approach into the **MiningTypestates** algorithm. It is easy to adapt the algorithm to use the approach that monitors all potential field objects. After all events in the trace have been processed, we union typestate models of all objects of the same type for each field to get a final typestate model for the field of the type. The union of two typestate models consists of the union of the states and the union of the transitions of these two models into one model.

This algorithm traverses the input trace *T* only once. If *T* has m_1 method entry events (and correspondingly m_1 method exit events) of composite objects, m_2 method entry events (and correspondingly m_2 method exit events) of field objects, and *n* field assignment events, the algorithm **MiningTypestates** has the time complexity of

$$m_2 + n + \sum_{i=1}^{i=m_1} 2 * \left| e_i.F \right|$$

ei.F is determined by the runtime behavior of the called method and can be approximated by the code of the called method. It is common that a class have s small number of fields. For example, the **BufferedOutputStream** has three fields. We can expect that | *ei.F* | is a small number. Assume that *p* is the largest of all | *ei.F* | with *i* ranging from 1 to m_1, them the time complexity is approximated as $m_2+n+2* p* m_1$. **MiningTypestates** has linear time complexity. In contrast, the commonly used *kTail* algorithm [16] and typical PFSA learners [17] have the running time quadratic and cubic to the length of the input trace, respectively.

IV. Preliminary Evaluation

We wrote an tracing agent using JVMTI (Java Virtual Machine Tool Interface). JVMTI is convenient to trace programs in many aspects such as that it is easy to access the call stack and that we can attach a unique tag to every object. We record four types of events: method entry event, method exit event, field modification event and field access event. We wrote a small Java program to transform the raw trace into the new one with the format defined in Section II.B. This is as easy as to collect all fields of the receiver object that are written (read)by a field modification (access) event appearing between the entry event and the exit event of the method invocation. These fields are attached to the method entry event. All field access events are removed from the original trace. The recorded field modification event is just the field assignment event in our trace definition.

We applied our approach to mine typestate models for classes from three system packages and their sub-packages of the Oracle Java JDK 6: java.lang, java.util, and java.io, totally 17 packages. Classes in these packages obey important API properties. They are widely used as experimental targets in the literature [8][17][15]. We configured the tracing agent to record events of objects of the types within the target packages. For *method entry* and *method exit events*, only these ones whose called methods are public constructors or public, instance methods are recorded. Training programs used in our experiments are 11 benchmarks from the DaCapo benchmark suite 2006-10-MR2, which ensures a controlled and reproducible execution of all benchmarks [23]. Considering numerous events produced, The execution time of every program is limited to at most two hours. The benchmark programs and their traces are presented in Table I. All experiments were carried out on a machine of Win7 and 4G RAM, 3.0 GHz Intel Core i5-2320 CPU with the 64- Bit Server VM of the Oracle Java SE 1.6.0_27. In the paper, we collected the performance data by repeating each run 10 times and the geometric means

TABLE I
Overview of benchamarks and experimental results.

Benchmark	Events	Models	Time (m)	Time Increase
antlr	57,538,679	80	6.26	12.28%
bloat	81,062,987	82	10.23	13.43%
chart	78,233,879	145	9.20	7.85%
eclipse	104,053,673	91	15.08	7.03%
fop	40,315,658	89	5.07	3.28%
hsqldb	77,076,574	82	8.27	10.26%
jython	59,835,239	77	19.23	5.15%
luindex	110,391,388	83	16.20	8.91%
lusearch	93,464,648	79	7.34	6.28%
pmd	52,876,129	92	14.28	5.91%
xalan	69,507,013	82	5.26	9.07%
total	824,355,867	154	116.42	8.13%

are computed as the result. In our empirical validation, we monitor all potential field objects.

We mine typestates for objects whose runtime types are of concrete and mutable classes (we ignore immutable objects such as objects of **String** and **Integer**) in the 17 target packages except local and anonymous classes.

For the input specifications of field objects, we use the specification of the interface **Closeable** presented in Figure 4. Despite limited expressive power of such same FSMs, they are important [4] and can describe important temporal properties. There are several techniques to mine them automatically [10][26]. The Java API documentation states that "a **Closeable** is a source or destination of data that can be closed" and that its only method **close** "closes this stream and releases any system resources associated with it". We can see that there are two important temporal properties of the **Closeable**. The first liveness property is the resource-releasing specification that a **Closeable** should be closed if it will not be used again. The second safety properties is that after a **Closeable** is closed, it can not be used again, which holds in most cases. In the target packages, there are 56 concrete types implementing this interface that are not local or anonymous classes, and there are 32 target concrete classes that have at least one field of the **Closeable** type. These considerable numbers manifest that our approach can be intensively experienced. If this specification can effectively enhance mined typestates, we have reasons to believe that our approach will perform well to distinguish states of composite objects when more complex specifications are provided an input. This well-known specification facilitates the experimental validation of our approach, considering that many large specifications automatically mined by existing

Fig. 4 Input specification of concrete classes of the Closeable type in our evaluation.

ISBN: 978-1-61804-236-1

approaches are neither complete nor precise. We wrote a small program to automatically generate this specification for the 56 implementing classes of the **Closeable** type using the Java reflection utilities.

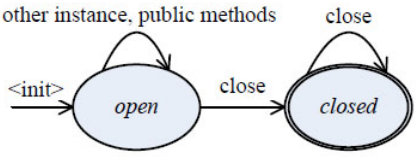


Table I presents the benchmark programs, the number of events for each benchmarks, the number of mined typestate models, the time cost to mine traces of each benchmark, and the time increase compared with the null-abstraction approach. The *total* cell of the *time increase* column is the average of the time increases of the 11 benchmark programs. Because different benchmarks may use objects of the same class, we mined several models for a class, each from a different benchmark program. In such cases, we unionized these models into one model as the final typestates for the class. In total, 154 typestate models are mined. Our approach is very fast. The analysis time is roughly linear to the length of the input trace. Mining one trace of several tens of millions of events typical takes around 10 minutes and none of the input traces exceeds 20 minutes. Compared with the null-abstraction approach, the extra execution time is always between 3-14%. This overhead is limited considering the improved precision obtained by our approach.

To validate mined specifications, we manually inspected mined models. The main reference of the inspection was the Java documentation and the source code of the target classes. Sub-models for fields that are not of the **Closeable** type mined by the null-abstraction approach and ours are same because the same state abstraction function is used for them. The comparison is between sub-models for fields of the **Closeable** type. The sub-models for such fields mined by the null-abstraction approach is similar to the one presented in Figure 1 (b) while the sub-models for such fields mined by our approach is similar to the one presented in Figure 1 (d).

Among the 32 target concrete types that have at least one field of the **Closeable** type. There are 17 ones to the fields of whose objects have been assigned some **Closeable** objects during runtime of the benchmark programs. We compared the models mined by the null-abstraction approach and those mined by our approach for these 17 classes to evaluate the effectiveness of our approach to enhance mined typestates. The results are presented in Table II. The initial models denote typestate models mined by the null-abstraction approach and the enhanced models denote typestate models mined by our approach. The initial models for 8 *out* of the 17 classes are correct in terms of the specification presented in Figure 4. The enhanced models for these classes are the same as the initial ones. Most of these 7 classes have some state-indicating fields such as the out field of the **BufferedWriter** that is initialized by the constructor and is set to be null by the close method. The sub-models for these state-indicating fields mined by the null-abstraction approach are the same as those for the wrapped streams mined by our approach. The **StringWriter** has an empty **close** method. The models for the left 9 classes are all overgeneralized in that further operations are permitted after their objects are closed. 6 out of the 9 overly general models are enhanced by our approach. Exactly one additional *closed* state is identified as the final state of the sub-models for the wrapped streams, which exactly reflects the specification of the **Closeable**. Figure 1 (d) presents one such enhanced model with uninteresting methods omitted. The left 3 overgeneralized models are not enhanced. The reason for the **OutputStreamWriter** and **PrintStream** is that we did not observe the close invocations in the traces. This is an inherent limitation of all dynamic specification mining approaches because we can only generalize based on observed behaviors. We can not enhance the model of the **InputStreamReader** because no specification is provided for its field sd of the type **sun.nio.cs.StreamDecoder**. The **close** method of the **InputStreamReader** calls the method *sd*.**close**, not directly calls the **close** of its wrapped input stream. In all, we successfully enhance 66% (6 out of 9) of the overgeneralized models.

TABLE II
MODEL ENHANCEMENT.

Subject	Initial Models	Enhanced Models	Enhancement
`BufferedInputStream`	correct	remain same	0
`BufferedWriter`			0
`PrintWriter`			0
`PushbackInputStream`			0
`StringReader`			0
`StringWriter`			0
`ZipInputStream`			0
`DeflaterOutputStream`			0
`BufferedOutputStream`	overgeneralized	enhanced. Exactly one additional *closed* states is identified as the destination state for each `close` transition.	75.02%
`BufferedReader`			73.21%
`DataInputStream`			68.90%
`DataOutputStream`			78.34%
`FileReader`			80.20%
`FileWriter`			76.28%
`OutputStreamWriter`	overgeneralized	remain same. No `close` invocations observed.	0
`PrintStream`			0
`InputStreamReader`	overgeneralized	remain same. No specification provided for field *sd*.	0

Although we can easily see that the quality of mined models is enhanced by our approach compared with those mined by the null-abstraction approach, we try to quantitatively evaluate the enhancement. In the literature, the measurement of *precision* (the percentage of mined behavior that is correct) and *recall* (the percentage of correct behavior that has been mined) are often used [16][10][9]. After the manual inspection, we observed that no recall is lost for all of the enhanced models

compared with their corresponding initial ones for the case of our benchmark programs. This resulted from the fact that our approach just added necessary states to models of composite objects to make them obey specifications of their field objects. Theoretically speaking, it is possible for our approach to add redundant states to models of composite objects and thus lead to recall loss. However, we did not observe this in our experiments.

Considering the fact that our approach did not lose recall for these models, we adopted a convenient way to evaluate the precision enhancement. We define the *precision enhancement* as the percentage of the behavior of the initial model that are rejected by the enhanced model. This rejected behavior is erroneous behavior that is incorporated in the initial model but removed from the enhanced model by our approach. To perform this evaluation, we applied the trace generation algorithm TraceGen [9] to randomly generate normal traces from the initial model and then simulated these traces on the corresponding enhanced model. A *normal trace* is a sequence of transition labels (methods) that form a path starting from the initial state to a final state of the FSM model. It represents normal behavior of the object that is accepted by its model. To generate a normal trace, we start from the initial state of the model and randomly choose an outgoing transition to reach the next state. This repeats until we reach a final state. The precision enhancement is the ratio of the number of traces that are rejected by the enhanced model to the number of all generated traces from the initial model.

For example, for the initial model presented in Figure 1 (b) of the **BufferedOutputStream**, we can generate a set of two normal traces that cover all of its transitions at least once:

T_1: <**init**>, **write**, **write**, **close**.

T_2: <**init**>, **write**, **close**, **write**.

We can see that T_1 represents normal behavior while T_2 represents erroneous behavior. Then we simulate these two trace on the enhanced model presented in Figure 1 (d). T_1 is accepted, while T_2 is rejected. So, for this set of normal traces, the precision enhancement is 1/2 = 50%.

Because sub-models for fields that are not of the **Closeable** type mined by the null-abstraction approach and ours are same, we compute the precision enhancement of sub-models for fields of the **Closeable** type mined by our approach over those mined by the null-abstraction approach. We configured the **TraceGen** algorithm to generate sets of normal traces such that each transition in the initial model had to be covered at least 3 times. We computed the precision enhancement by repeating each experiment 10 times with 10 different sets of normal traces, and the average of the results is presented in the *Enhancement* column of Table I. The average precision enhancement for these six models is 75.33%.

A. Threats to Validity

We consider all composite classes from 17 packages of the Java system library and use the DaCapo suite as the training programs. We only use one simple property of the **Closeable** as input. Although in target packages there are 56 concrete types implementing this interface and 32 target concrete classes that have at least one **Closeable** field, the composite classes and field classes that experience our approach are limited: the 17 composite classes comes from two packages and are all streams of different kinds: **java.io** and **java.util.zip**. Further studies with subjects from other libraries and other domains are useful. Our process of manually inspecting mined typestates may be subject to errors. We therefore employed two inspectors to repeat this work and compared their results to avoid unintentional mistakes Each programmer inspected all these models independently and at last compared the inspection results. If there were some conflicts, they performed further inspection to get the coincidence.

V. Related Work

Liblit et al. [27] propose to use predicates on variables to capture program behavior to isolate bugs of C programs. The used predicates include six relations >, ⩾, <, ⩽, = and ≠ between the return of a scalar-returning function and 0, and between a scalar variable and another in-scope scalar variable or an in-scope scalar constant expression. Inspired by this work, Dallmeier et al. [7], [15] propose to mine typestates of objects by utilizing an abstraction on values of object fields (or returns of observer methods [7]). The abstraction maps concrete values of object fields to abstract values as follows: values of reference types and arrays are abstracted to *null* (=*null*) or *not null* (≠ *null*), values of numerical fields are abstracted to *larger than zero* (>0), *less than zero* (<0), or *equal to zero* (=0), and values of boolean fields remain unchanged. The set of abstract values of all fields of an object identifies an abstract state of the object. They report that this abstraction strikes a good balance between the expressiveness of the model and its size.

Dallmeier et al. [22] propose to mine *deep models* that also include states of transitively reachable objects rather than just a ≠*null* state for a field object of reference type in the common *shallow* models. A *depth* parameter is introduced to define how many times we consider states of deeper reference fields when constructing abstract states. The common models have the depth of 0 and a model with the depth of 1 also considers the values of fields of reference fields of the target object. Reference fields beyond the depth are still abstracted to =*null* and ≠*null*. To discover all states, this approach depends on the existence of state-indicating fields in the composite object or in its descendant field objects, which can not be guaranteed. In addition, considering unrelated fields can lead to unnecessary states that complicate mined typestates. In general, we do not know what is the best abstraction for specific classes without further knowledge. We believe that there is not a best and general abstraction even for the primitive numeric types. Simply transitively abstract the fields of a reference field may be not appropriate. Moreover, to keep the approach to be scalable, the depth of the model must be small (Dallmeier et al. [22] mines models with the depth of 1), and choosing an appropriate depth for the general typestate mining is difficult. In contrast, if there are already reliable specifications for afield object, we can utilize it.

Ghezzi et al. [19] present the SPY approach to mine

typestates of Java data container classes. The return values of inspector methods are used to label states. They consider returns and parameters of method invocations and their concrete values are used. The typestates mined can be very precise but at the same time can be extremely large and complex. In fact, a finite number of FSMs can not describe their mined models. So, they build a set of graph transformation rules to generate these models intensionally. Compared with SPY, other approaches [10] including ours use abstract field values to label states and produce models of moderate size with balanced precision and recall. Furthermore, the complexity of mined typestates by our approach depends on and can be controlled by the input specifications. Simple input specifications tendto produce less complex typestates. SPY do not rely on the fields of target classes, while our approach need to inspect the internal states of objects.

Lo et al. [16] present a steering mechanism to improve the precision of specifications mined by the *kTail* algorithm. They first mine execution traces to infer simple temporal properties. These properties have the potential to capture constraints of distant events. Then, they use the *kTail* algorithm to mine specifications. The inferred temporal properties are used to guide the merging of equivalent states of the *kTail* algorithm. Two equivalent states are merged only if the merging does violate any temporal property. This approach is automatabased, and it does not consider the explicit states of programs. The inferred simple properties may capture some constraints of subcomponents, but more benefits of available specifications of subcomponents are not explored.

Lo and Maoz [28] propose to mine statistically significant LSCs (live sequence charts) at the user-defined abstraction level. The mined LSCs can be refined to lower level LSCs or abstracted to higher level LSCs at the user's requests. These mined LSCs are temporal invariants on two sequences of inter-object method calls. Because the concept of inter-object depends on the chosen abstraction level, high abstraction levels result in good scalability. The Java package hierarchy of the inclusion relation between packages and classes is used to provide various abstraction levels. Different from their approach, we mine typestates of single objects and leverage the composition relation between objects to improve the precision of mined typestates.

VI. Conclusions

This paper presents a framework to mine precise typestates of composite objects by leveraging the specifications of field objects. To cope with the overgeneralization introduced by the null-abstraction approach, instead of the single $\neq$*null* state, we distinguish different states of field objects encoded in their available specifications and use them to construct abstract states of the composite object. As a result, the states of the field objects become visible to the composite object. Many new states of the composite object can be found and the precision of mined typestates of the composite objects is improved. In future work, we plan to conduct further empirical studies with subjects from other libraries. In addition, we plan to explore the possibility of leveraging input specifications of other forms, such as the invariants over variables.

References

[1] R. M. Hierons, K. Bogdanov, J. P. Bowen, R. Cleaveland, J. Derrick,J. Dick, M. Gheorghe, M. Harman, K. Kapoor, P. Krause, G. L¨uttgen, A. J. H. Simons, S. Vilkomir, M. R. Woodward, and H. Zedan "Using formal specifications to support testing," *ACM Comput. Surv.*, vol. 41, no. 2, pp. 9:1–9:76, 2009.

[2] M. Pradel, C. Jaspan, J. Aldrich, and T. R. Gross, "Statically checking API protocol conformance with mined multi-object specifications," in *ICSE*, pp. 925–935, 2012.

[3] M. Pradel and T. R. Gross, "Leveraging test generation and specification mining for automated bug detection without false positives," in *ICSE*, pp. 288–298, 2012.

[4] W. Weimer and G. Necula, "Mining temporal specifications for error detection," in *TACAS*, pp. 461–476, 2005.

[5] G. Ammons, R. Bod´ık, and J. R. Larus, "Mining specifications," in *POPL*, pp. 4–16, 2002.

[6] J. Whaley, M. C. Martin, and M. S. Lam, "Automatic extraction of object-oriented component interfaces," in *ISSTA*, pp. 218–228, 2002.

[7] V. Dallmeier, C. Lindig, A. Wasylkowski, and A. Zeller, "Mining object behavior with adabu," in *WODA*, pp. 17–24, 2006.

[8] M. Pradel and T. R. Gross, "Automatic Generation of Object Usage Specifications from Large Method Traces," in *ASE*, pp. 371–382, 2009.

[9] D. Lo and S.-C. Khoo, "Quark: Empirical assessment of automatonbased specification miners," in *WCRE*, pp. 51–60, 2006.

[10] C. Le Goues and W. Weimer, "Measuring code quality to improve specification mining," IEEE Transactions on Software Engineering, vol. 38, no. 1, pp. 175–190, 2012.

[11] M. Gabel and Z. Su, "Testing mined specifications," in *FSE*, pp. 4:1– 4:11, 2012.

[12] R. Strom and S. Yemini, "Typestate: A programming language concept for enhancing software reliability," *IEEE Transactions on Software Engineering*, vol. SE-12, no. 1, pp. 157–171, 1986.

[13] A. Mesbah and A. van Deursen, "Invariant-based automatic testing of AJAX user interfaces," in *ICSE*, pp. 210–220, 2009.

[14] L. Mariani, A. Marchetto, C. D. Nguyen, P. Tonella, and A. Baars, "Revolution: automatic evolution of mined specifications," in *ISSRE*, pp. 241–250, 2012.

[15] V. Dallmeier, N. Knopp, C. Mallon, G. Fraser, S. Hack, and A. Zeller, "Automatically generating test cases for specification mining," IEEE Transactions on Software Engineering, vol. 38, no. 2, pp. 243–257, 2012.

[16] D. Lo, L. Mariani, and M. Pezze, "Automatic steering of behavioral model inference," in *ESEC/FSE*, pp. 345–354, 2009.

[17] C. Lee, F. Chen, and G. Ros¸u, "Mining parametric specifications," in *ICSE*, pp. 591–600, 2011.

[18] A.W. Biermann and J. A. Feldman, "On the synthesis of finitestate machines from samples of their behavior," IEEE Transactions on Computers, vol. C-21, no. 6, pp. 592–597, 1972.

[19] C.Ghezzi, A. Mocci, and M. Monga, "Synthesizing intensional behavior models by graph transformation," in *ICSE*, pp. 430–440, 2009.

[20] M. Robillard, E. Bodden, D. Kawrykow, M. Mezini, and T. Ratchford, "Automated API property inference techniques," *IEEE Transactions on Software Engineering*, vol. 99, no. 1, 2012.

[21] M. D. Ernst, J. H. Perkins, P. J. Guo, S. McCamant, C. Pacheco, M. S. Tschantz, and C. Xiao, "The Daikon system for dynamic detection of likely invariants," *Science of Computer Programming*, vol. 69, pp. 35–45, 2007.

[22] V. Dallmeier, A. Zeller, and B. Meyer, "Generating fixes from object behavior anomalies," in *ASE*, pp. 550–554, 2009.

[23] S. M. Blackburn, R. Garner, C. Ho_mann, A. M. Khang, K. S. McKinley, R. Bentzur, A. Diwan, D. Feinberg, D. Frampton, S. Z. Guyer, M. Hirzel, A. Hosking, M. Jump, H. Lee, J. E. B. Moss, A. Phansalkar, D. Stefanovi´c, T. VanDrunen, D. von Dincklage, and B. Wiedermann, "The DaCapo benchmarks: java benchmarking development and analysis," in *OOPSLA*, pp. 169–190, 2006.

[24] R. DeLine and M. Fahndrich, "Typestates for objects," in *Proceedings of the European Conference on Object-Oriented Programming (ECOOP 2004)*, pp. 465-490, 2004.

[25] K. Bierho_ and J. Aldrich, "Lightweight object specification with

typestates," in *Proceedings of the 10th European software engineering conference held jointly with 13th ACM SIGSOFT international symposium on Foundations of software engineering (ESEC/FSE-13)*, pp. 217- 226, 2005.

[26] Q. Wu, G. Liang, Q. Wang, T. Xie, and H. Mei, "Iterative mining of resource-releasing specifications," in *ASE*, pp. 233–242, 2011.

[27] B.Liblit, M. Naik, A. X. Zheng, A. Aiken, and M. I. Jordan, "Scalable statistical bug isolation,"in Proceedings of the 2005 ACM SIGPLAN conference on Programming language design and implementation (PLDI), pp. 15–26, 2005.

[28] D.Lo and S. Maoz, "Mining hierarchical scenario-based specifications," in Proceedings of the 2009 IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 359–370, 2009.

Fuzzy Ontology-Based Model for Information Retrieval

Zeinab E. Attia

***Abstract*—** The paper proposes a linguistic-based information retrieval model. It has a linguistic-based query answering system to deal with a user linguistic-based queries in a certain field (view). Using these linguistic-based queries, users has the ability to define their needs accurately. The model also has the ability to deal with the multi-field topics problem using a predefined multi-field or multi-view fuzzy ontology. The model also enhances the recall measure respecting another two fuzzy ontology-based information retrieval models. The model also proposes a ranking algorithm that ranks the set of relevant documents according to some criteria such as their relevance degree, confidence degree, and updating degree.

***Keywords*—** Information Retrieval, Fuzzy ontology-based information retrieval, fuzzy ontology.

I. Introduction

An information retrieval system (IR) consists of a document collection, a user query, a retrieval engine, and a ranking module. It stores and annotates documents such that when users express their information needs in a query, the ranking module will show a set of ranked relevant documents. This set of documents is retrieved by the retrieval engine associating a score to each one. The higher the score is, the greater the document relevance [8]. So, the challenge in IR is to find a number of the most relevant documents according to user's query.

Researchers deal with this challenge using two different approaches. These approaches are keyword based approach and concept based approach. In the keyword based approach, documents are returned when they are annotated by terms specified in the searching query. However, this approach neglects many related documents that are not annotated with the query terms [8]. In the concept based approach, documents are returned according to their relevance to the searching query. This approach is a domain specific approach. It can be classified into ontology based approach and fuzzy ontology based approach. The performance of any IR system is measured using many computing parameters which are recall, precision, fmeasure… and many more [11].

Unfortunately, the current information retrieval systems suffer from many problems. Some of them are low in precision, low in recall, inability to deal with the multi-field topics problem and inability to allow their users to define their needs accurately.

Recall is the proportional of the correctly retrieved documents among the pertinent documents in the collection [12]. Precision is the proportion of the correctly retrieved documents among the documents retrieved by the system [12]. Multi-field topics are topics that combine two or more fields together such as the "bioinformatics" that combines the medical field with the computer science field. When certain medical user searches for a bioinformatics paper, the IR system will return the same set of documents that are returned to a computer science user. So these systems do not have the ability to distinguish between results of such topics respecting the field point of view.

The paper proposes a linguistic-based Fuzzy Ontology Information Retrieval model. It has the ability to deal with the multi-field topics problem, allow its users to define their needs accurately. Also it aims to increase the recall measure respecting Leite [4] and FROM [5] models.

The rest of the paper is organized as follow; section II presents fuzzy ontology. Fuzzy ontology based Information Retrieval is discussed in section III. Section IV shows some related work. The proposed Linguistic based Fuzzy Ontology Information Retrieval model is presented in section V. Section VI shows a case study to test the proposed model. The paper is concluded in section VII.

II. Fuzzy Ontology

Ontology is "*the conceptualization of a domain into a human understandable, machine readable format consisting of entities, attributes, relationships, and axioms*". It is used as a standard knowledge representation for the semantic web [2]. Unfortunately, the conceptual formalism, supported by typical ontology, may not be sufficient to represent uncertain information commonly found in many application domains. This is due to the lack of clear-cut boundaries between concepts of the domains. Moreover, fuzzy knowledge plays an important role in many domains that face a huge amount of imprecise and vague knowledge and information, such as text

This work was supported in part by the U.S. Department of Commerce under Grant BS123456 (sponsor and financial support acknowledgment goes here). Paper titles should be written in uppercase and lowercase letters, not all uppercase. Avoid writing long formulas with subscripts in the title; short formulas that identify the elements are fine (e.g., "Nd–Fe–B"). Do not write "(Invited)" in the title. Full names of authors are preferred in the author field, but are not required. Put a space between authors' initials.

F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (corresponding author to provide phone: 303-555-5555; fax: 303-555-5555; e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar. colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

mining, multimedia information system, medical informatics, machine learning, and human natural language processing. To handle uncertainty of information and knowledge, one possible solution is to incorporate fuzzy theory into ontology [9] yielding a fuzzy ontology model.

Accordingly, fuzzy ontologies will contain fuzzy concepts and fuzzy memberships. Fuzzy ontologies are capable of dealing with fuzzy knowledge, and are efficient in text and multimedia object representation and retrieval [3]. There are many fuzzy ontology definitions according to the underlined application and domain. Some of them are:

[1] fuzzy ontology is a pair (C, R), where C is a set of concepts, R is a set of fuzzy relations between concepts.

[10; 11] fuzzy ontology is a quadruple(C, R, P, I), where C is a set of fuzzy concepts, R is a set of binary relations, P is a set of fuzzy properties of concepts, I is a set of individuals.

[14; 15] fuzzy ontology is a quadruple(C, R, F, U), where C is a set of concepts, R is a set of fuzzy abstract relations, F is a set of fuzzy concrete relations, U is the universe of discourse.

III. Fuzzy Ontology-based information retrieval

Fuzzy Ontology based Information Retrieval model, FOIR, is an IR model that semantically retrieves a set of relevant documents with respect to a certain query in a specific domain. This domain is represented using fuzzy ontology [5, 7, 8].Commonly, FOIR has three main components including input, retrieval processing, and output components. The input component includes document collection, fuzzy ontology, and user's query. Retrieval engine and ranking module are retrieval processing components. The output component is the set of resulted ranked relevant documents. FOIR has four phases which are: document annotation, query expansion, retrieving a set of relevant documents, ranking a set of relevant documents.

FOIR takes as input a set of documents, and a user query, to retrieve a set of the most relevant documents with respect to the entered query using a retrieval engine, then ranks this set and return it to the user. Both the document annotation process and the query expansion process depend on a fuzzy ontology.

IV. Related work

Leite model [4] semantically retrieves a set of query's relevant documents in multi-domains. Each domain is represented as a fuzzy ontology and is then connected with other domains using fuzzy positive relations. It uses the well known "tfidf" method to annotate the document collection with a set of fuzzy ontology concepts. It deals with crisp queries. When a certain user enters a query, Leite expands it using a two phases query expansion process. The first phase expands each concept in the query with all of its related concepts in other domains. Then the result enters the second phase to expand each concept in it with all of its related concepts in the same domain. The max product composition between each document and the expanded user query is used as the similarity function to determine a set of the most relevant documents. This set of relevant documents is ranked in a descending order according to their relevance degree and returned to the user.

Fuzzy Relational Ontology Model, FROM, [5] is a document retrieval model based on fuzzy ontology. It semantically retrieves a set of relevant documents with respect to a user query. It assumes each document in the document collection is already annotated with a set of weighted keywords. It considers fuzzy ontology as a set of concepts, terms, and relations between concepts and terms. FROM deals with crisp queries. When a user enters his query, it expands each concept in it with all terms that describes it and each term in it with all concepts that it describes. It retrieves a set of relevant documents using the max min composition between each document in the document collection and the expanded user query. The resulted set is ranked in a descending order according to each document relevance degree and then it returned to the user.

Fernández model [6] proposed an ontology based information retrieval model. This model deals with open environment. It annotates the document collection using two techniques. The first one is an NLP based, while the second is a context semantic information based. When a certain user enters a query, the model performs some processing on it using the ontology-based Question Answering (QA) system, PowerAqua. The adaptation of the traditional vector space IR model is used as to calculate the relevance degree of each document in the document collection with respect to the entered user query. Documents are returned to the user such that documents with higher relevance degree are listed first.

Ranked Neuro Fuzzy Inference System, RNFIS, [7] proposed a hybrid information retrieval model. It is based on fuzzy version of vector space for information retrieval and fuzzy enhanced Boolean theory for document scoring. The model divides each document in the document collection into different weighted zones. When a certain user enters a query, the model expands it with multiple synonym queries based on its semantics. Then, it calculates the parameters; term frequency (tf), inverse term frequency (idf), and overlap (number of query terms found in the document) for each term of the expanded query in each zone in the document. These parameters in addition to the zone weight are fuzzified using the gaussian membership function. The retrieval engine uses these fuzzifed parameters in Sugeno's fuzzy rules. Then, it uses a certain aggregation operation to combine the result of all these rules. The result is then defuzzified to determine a crisp value. This result represents the relevance degree of this document respecting the user query.

All of these models can only deal with crisp queries. They are not able to deal with linguistic-based ones. So, its users can not define their needs accurately. Also, they suffer from low in the recall measure, as the result of using incomplete fuzzy ontology components for expanding a certain user query keywords. Also, they cannot handle the multi-field topics problem. To rank the resulted documents, these models use the similarity degree between each document in the document collection and the user query keywords.

V. The proposed Linguistic based fuzzy ontology information retrieval model

The proposed model is a linguistic-based semantic document retrieval model that uses a predefined multi-view fuzzy

ontology. It semantically retrieves relevant documents according to a user's linguistic based query. It can be used to retrieve any kind of documents in a specific domain written in any language. The proposed model main features are:

- Increase the recall measure respecting FROM [5] and Leite [4] IR models. As its expansion algorithm uses a fuzzy ontology with components a set of concepts, relation between them, terms, relation between them, and a set of relations between concepts and terms.
- Deal with the multi-field topics problem. This is through using a predefined multi-view fuzzy ontology during its expansion algorithm to expand each user keyword in a certain field or view.
- Allow its users to define their needs accurately using linguistic-based queries, e.g., select all papers that are *very related* to *bioinformatics* in the *computer science* field search. "*very related*" is a linguistic term, "*bioinformatics*" is a keyword, "*computer science*" is the field or view point of search.
- Rank the resulted semantically relevant documents according to some criteria, such as the document matching degree, its confidence degree, and its timeliness.

A. *The proposed Information Retrieval Structure*

The proposed information retrieval model's main components are a set of annotated documents, users' profiles, users' queries, retrieval engine, and ranking module. It depends mainly on fuzzy ontology methodology and some NLP tools such as stemmer, POS tool.

Figure1 shows the structure of the proposed model. Firstly, each user should create a profile to define all his linguistic terms. Now, the user can build his query. This query is a set of keywords each is associated with its importance degree. This importance degree is expressed in linguistic terms. For example, select all papers that are *very relevant* to *bioinformatics* in *computer science* search point of view, here the user searches for papers that are *very related* (linguistic term) to the keyword *bioinformatics* (keyword) according to the *computer science* search point of view. This query is then passed on some operations, which are:

- Interpreting each linguistic term according to the user's subjective view,
- Expanding each keyword with its related keywords using the predefined fuzzy ontology in its specified search point of view.

Then, this expanded list enters the retrieval phase that semantically retrieves a set of matched documents each associated with a matching degree. This set is then ranked according to some criteria using the proposed ranking algorithm. Finally, the ranked relevant set of documents is displayed to the user.

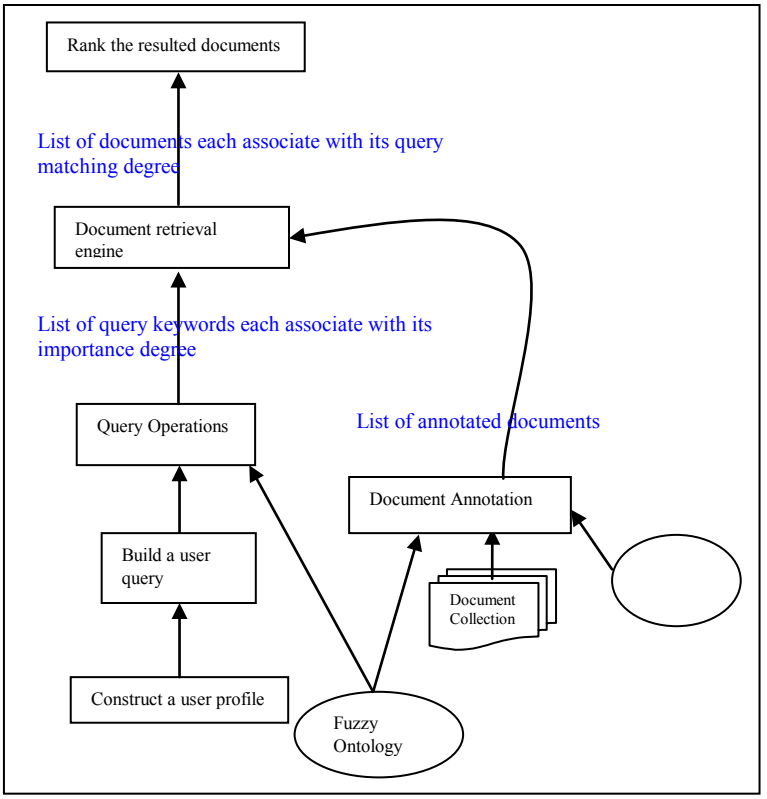


Figure 1: The proposed model phases

B. *The proposed Fuzzy Ontology Tool*

The proposed fuzzy ontology model is a Multi-Views Fuzzy Related Ontologies, MVFRO [10]. Some of its main features are listed below:

- It is a general multi-domain fuzzy ontology, which can fit any domain and any application.
- The main fuzzy ontology components are concepts, relations, properties, terms, and individuals.
- The relation between fuzzy ontology components or the related fuzzy ontologies can have multi-fuzzy-values each represents a certain point of view, e.g., In the old English, poetry represents the English literature with degree about 0.3, while in the modern English, poetry represents about 0.25 from the English literature.
- Using linguistic values and fuzzy number to express the relation between fuzzy ontology components or the relation between the related fuzzy ontologies.
- The used linguistic values and fuzzy numbers are defined by the domain expert according to his own subjective view.
- Storing all ontology components after stemming it in a relational database.
- Sorting different point-of-views that represent a certain relation between the ontology components or the related ontologies in one table instead of having one table per view.

ISBN: 978-1-61804-236-1

- Storing the expert's subjective view about each used fuzzy number and linguistic term.

C. *proposed model phases*

The proposed model phases are as follows:

1. User profile creation

User should create an account before building his/her query. Using this account, he/she can define any linguistic term according to his subjective view. Figure 2 shows the scheme of storing users' linguistic terms definitions. When a certain user creates an account, this account is stored in the *Users* table. Then user can define any linguistic term, e.g., "related" is a linguistic term, using the userLinguisticTermFunction table. This table specifies which membership function is used to define a certain linguistic term according to the user subjective view. This membership function is also defined according to the user's subjective view and stored in a table correspond to its name, e.g., *triangularUserLinguisticTerms* table for triangular membership function, *piUserLinguisticTerms* table for pi membership function. Hedges, e.g., very, more or less are hedges, can also be defined according to user's subjective view and stored in *userHedgeDefinition* table. In *userHedgeDefinition* table user specifies the hedge name and its power. Also, users have the ability to specify which method is used to interpret a conjunctive or disjunctive query, by determining the conjunctive and disjunctive methods and storing them in *userConjunctiveDisjunctiveMethod* table.

Now, user can build his query. Some query operations are then performed on this query.

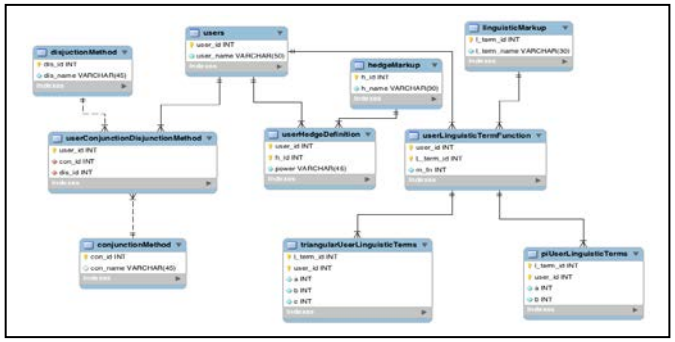


Figure 2: storing a user linguistic terms definition

2. Constructing a linguistic based multi-view query

Now, the user can build his query. This query can be crisp, fuzzy, or linguistic based-query. For example, the query statement, "select all data mining papers" represents a crisp query. On the other hand, the query statement, "select all data mining papers with membership degree 0.6" and "select all data mining papers with membership degree around 0.6" are examples of the fuzzy query. A linguistic based query may be like:

select all papers very related to bioinformatics according to the medical view

where "*bioinformatics*" is the keyword that the user searches for. "*very related*" is a user linguistic term that reflects his needs for the keyword "*bioinformatics*". "*medical*" is the search point of view. This linguistic term is previously defined by the user according to his subjective view and stored in his account.

3. Applying the Query Operations

After user submits his query, some operations are performed on it. First the query is parsed, such that each searched keyword is extracted with its importance degree that is expressed using linguistic terms and hedges and with the search point of view. All linguistic terms and hedges are then interpreted according to the user's subjective view. Each keyword is then expanded in its specified search point of view using the domain fuzzy ontology. The importance degree of any expanded word is the product of its relation with the original keyword and the importance degree of the original keyword.

4. Retrieving a set of relevant documents

It semantically retrieves a set of relevant documents with respect to a certain user query through calculating document matching degree. A document matching degree is calculated as the max product composition between the list of weighted keywords that annotate this document and the list of query's weighted expanded keywords.

The result of this is a list of semantically relevant documents each associated with its matching degree.

5. Ranking the resulted documents

It ranks the resulted semantically relevant documents from the retrieval phase based on some criteria:

- The document's matching degree with user needs. The higher the matching degree is, the more document relevance with respect to user's needs.
- The document's confidence degree. This degree is extracted from the document's authors, the confidence degree of the journal, or conference that the document is published in. This factor reflects to what extent does the knowledge in this document is trusted. The higher the journal impact degree is, the more confidence that the knowledge in this document is correct,
- The document's updating degree. This degree is extracted from the document publishing date. This factor reflects to what extent does the knowledge in this document is new and updated, not out of date.

The ranked list of relevant documents is then displayed to the user in the same order.

VI. APPLYING OUR PROPOSED MODEL ON FROM CASE STUDY

This section tests the proposed model on FROM case study [5]. Figure 3 shows some changes in FROM fuzzy ontology. Considering fuzzy ontology, it represents the computational intellegence domain in the theoritical point of view. Regarding fuzzy ontology structure, it also includes a set of relations between concepts and each other. All relations are represented as fuzzy numbers instead of membership degrees, for more realistic and accuracy in describing this relations. Consider the fuzzy number 'around' is defined by the expert using the triangular membership whose parameters 'a' and 'c' have the values '-0.1' and '+0.1' respectively. All the fuzzy ontology relations are interpreted using this definition then stored in the

proposed model's database as in figure 4. Since the ontology size is small, we choose inferring any new relation during its insertion time and store them into the database. This will decrease any ontology query response time.

Regarding FROM case study document collection, we assume each is annotated with a set of weighted keywords, a string of its authors, its published date, the conference or the journal that publishes it. Considering the set of weighted keywords, we will deal with the same set that FROM case study works on. For other annotations, we assume their values and store them into the document annotation database. Figure5stores the document collection annotations in the database. For each document we store its annotated weighted terms, weighted concepts, its publishing date, its authors, and journal or conference of publishing it.

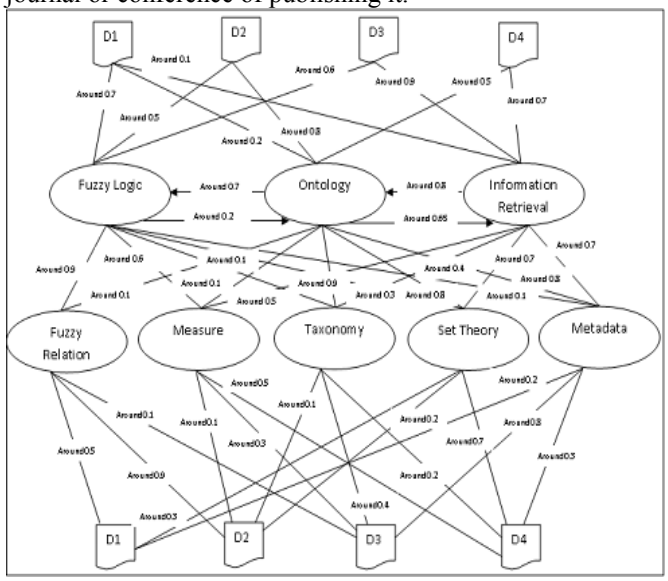


Fig. 3: applying the proposed model on FORM case

Let's consider the following linguistic based query, Q:

Q: ***"Ontology very related" OR "Fuzzy Relation related in the theoritical point of view"***

First, interpret each linguistic term in Q using the user definition:

Related = 0.88, *Very related* = 0.79

So, user query can be rewritten as

"Ontology 0.79" OR "Fuzzy Relation 0.88 in the theoritical view"

Second, expand the user query as follow:

1-Check the first keyword type whether it is represented in the ontology as a term or a concept:

"***Ontology***" is a concept

2-expand the concept "***ontology***" in the theoritcal of view as follow:

i- using its related concepts with degree multiplied by 0.79>= 0.6

{(Information Retrieval, 0.632)}.

ii- using terms that describes it with degree multiplied by 0.79 >= 0.65

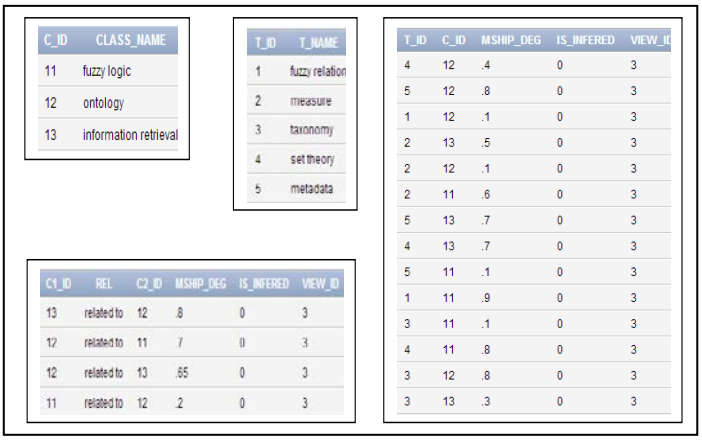
{(Taxonomy, 0.711), (Set Theory, 0.632)}.

iii- use the union operator between step 'i' and step 'ii' to have the expanded ontology set:

{(Information Retrieval, 0.632), {(set Theory, 0.632), (Taxonomy, 0.711)}.

iv- add the concept Ontology with degree 0.79 to step 'iii' to have the expanded ontology set:

{(Ontology, 0.79), (Information Retrieval, 0.632), {(set Theory, 0.632), (Taxonomy, 0.711)}.



C_ID	CLASS_NAME
11	fuzzy logic
12	ontology
13	information retrieval

T_ID	T_NAME
1	fuzzy relation
2	measure
3	taxonomy
4	set theory
5	metadata

C1_ID	REL	C2_ID	MSHIP_DEG	IS_INFERED	VIEW_ID
13	related to	12	.8	0	3
12	related to	11	.7	0	3
12	related to	13	.65	0	3
11	related to	12	.2	0	3

T_ID	C_ID	MSHIP_DEG	IS_INFERED	VIEW_ID
4	12	.4	0	3
5	12	.8	0	3
1	12	.1	0	3
2	13	.5	0	3
2	12	.1	0	3
2	11	.6	0	3
5	13	.7	0	3
4	13	.7	0	3
5	11	.1	0	3
1	11	.9	0	3
3	11	.1	0	3
4	11	.8	0	3
3	12	.8	0	3
3	13	.3	0	3

Fig. 4: storing FROM case study in database

3-Check the second keyword type whether it is represented in the fuzzy ontology as a term or a concept:

"***Fuzzy Relation***" is a term

4-expand the concept "***Fuzzy Relation*** " in the theoritical point of view as follow:

i-using its related concepts that it describes with degree multiplied by 0.88 >= 0.65

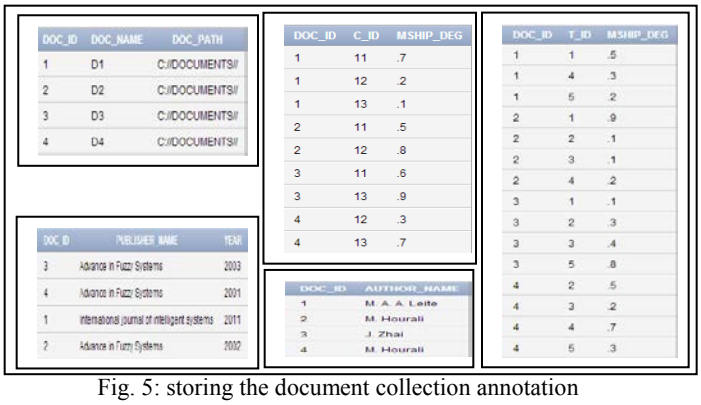
{(Fuzzy Logic, 0.792)}.

ii-add the term fuzzy relation with degree 0.88 to step 'i' to have the expanded fuzzy relation set:

{(fuzzy relation, 0.88), (Fuzzy Logic, 0.792)}.

5-Apply the union operator on the expanded ontology set and the expanded fuzzy relation set:

{(Ontology, 0.79), (Fuzzy Logic, 0.792), (Information Retrieval, 0.632), (set theory, 0.632), (taxonomy, 0.711), (fuzzy relation, 0.88)}.



DOC_ID	DOC_NAME	DOC_PATH
1	D1	C:\DOCUMENTS\
2	D2	C:\DOCUMENTS\
3	D3	C:\DOCUMENTS\
4	D4	C:\DOCUMENTS\

DOC_ID	C_ID	MSHIP_DEG
1	11	.7
1	12	.2
1	13	.1
2	11	.5
2	12	.8
3	11	.6
3	13	.9
4	12	.3
4	13	.7

DOC_ID	T_ID	MSHIP_DEG
1	1	.5
1	4	.3
1	5	.2
2	1	.9
2	2	.1
2	3	.1
2	4	.2
3	1	.1
3	2	.3
3	3	.4
3	5	.8
4	2	.5
4	3	.2
4	4	.7
4	5	.3

DOC_ID	PUBLISHER NAME	YEAR
3	Advance in Fuzzy Systems	2003
4	Advance in Fuzzy Systems	2001
1	International journal of intelligent systems	2011
2	Advance in Fuzzy Systems	2002

DOC_ID	AUTHOR_NAME
1	M. A. A. Leite
2	M. Hourali
3	J. Zhai
4	M. Hourali

Fig. 5: storing the document collection annotation in the a document annotation table

6-Divide the resulted expanded set into two sets, one for concepts and the other for terms:

Concept set= {(Information Retrieval, 0.632), (Ontology, 0.79), (Fuzzy Logic, 0.792)},

Term set= {(Set theory, 0.632), (Taxonomy, 0.711), (Fuzzy relation, 0.88)}.

Table 1: shows a list of journals each with its confidence degree

Journal name	**Weight**
International journal of intelligent systems	0.9
Knowledge and Information system	0.75
Advance in Fuzzy Systems	0.4

Third, use retrieval engine to retrieve a set of relevant documents, each with its relevancy degree:

For each of the four documents,

7-Calculate the max product composition for the document concept set with the query concept set:

R_C= {(D1, 0.5544), (D2, 0.632), (D3, 0.5688), (D4, 0.4424)}

8-Calculate the max product composition for the document term set with the query term set.

R_t = {(D1, 0.44), (D2, 0.792), (D3, 0.2844), (D4, 0.4424)}

9-Perform union operation on Rt and Rc:

R= {(D1, 0.5544), (D2, 0.792), (D3, 0.5688), (D4, 0.4424)}

10-Apply the threshold on the resulted relevant document set with value 0.4:

R= {(D1, 0.5544), (D2, 0.792), (D3, 0.5688), (D4, 0.4424)}

Table2 : shows a list of authors and their confidence degree

Author name	**Weight**
M. A. A. Leite	0.8
J. Zhai	0.7
M. Hourali	0.3

Fourth, apply the proposed ranking algorithm:

For each document:

11-Calculate its confidence degree weight, assuming table 1 and table 2, using Eq. 1:

$D_{Conf.Deg}$=0.3*max(journal_weight,author_weight) (1)

$D1_{Conf.\ Deg}$ *= 0.27,* $D2_{Conf.\ Deg}$ *=0.12,* $D3_{Conf.\ Deg}$ *=0.21,* $D4_{Conf.\ Deg}$ *=0.12}*

12-Calculate its updatence degree, using Eq. 2:

$D_{update.Deg}$=0.3*date_weight (2)

$D1_{update.\ Deg}$ *= 0.27,* $D2_{update.\ Deg}$ *=0.12,* $D3_{update.\ Deg}$ *=0.21,* $D4_{update.\ Deg}$ *=0.12}*

13-Calculate the document final weight, using Eq. 3:

D_{weight}=0.4*relevance_degree+$D_{Conf.Deg}$+$D_{updateDeg}$ (3)

Relevance list= {***D1= 0.76176, D2 = 0.5568, D3= 0.64752, D4= 0.41696***}.

14-Rank the relevance list in a descending order as follow:

The resulted relevance document= (***D1, D3, D2, D4***)

As we can see, adding the relation between concepts and each other return document D2 as it is about fuzzy logic which is related to ontology.

VII. Conclusion and Future work

This work presents an improvement in the fuzzy semantic information retrieval through:

- Building multi-views Linguistic based query system. This gives users more flexibility while building their queries.
- Allow users to define all their linguistic terms according to their subjective view. This helps in retrieving documents according to their linguistic terms definitions not to our definitions.
- Retrieve a set of relevant documents semantically using the proposed fuzzy ontology tool MVFRO.
- Deal with the multi-field topics problem using a predefined mutli-view fuzz ontology.
- The resulted set of documents is ranked according to some criteria which are their relevance degree with respect to use's query, confidence degree and updating degree.

The future direction to work in this area would be to build a document annotation algorithm using our proposed fuzzy ontology tool.

References

[1] L. Dey, M. Abulaish, "fuzzy ontologies for handling uncertainties and inconsistencies in domain knowledge description," the 17th IEEE international conference on fuzzy systems, 2008.

[2] Q. T. Tho, S. C. Hui, A. C. M. Fong, T. H. Cao," Automatic Fuzzy Ontology Generation for Semantic Web," IEEE transaction on knowledge and data engineering, Vol. 18, No.6, June 2006.

[3] J. Zhai, Y. Liang, Y. Yu, J. Jiang, "Semantic Information Retrieval Based on Fuzzy Ontology for Electronic Commerce," JOURNAL OF SOFTWARE, VOL. 3, NO. 9, DECEMBER 2008.

[4] M. A. A. Leite, I. L. M. Ricarte, "Relating ontologies with a fuzzy information model," KnowlInfSyst, pp. 619-651, 2013.

[5] R. Pereira, I. Ricarte, F. Gomide, " Information Retrieval with FROM: The Fuzzy Relational Ontological Model," INTERNATIONAL JOURNAL OF INTELLEGENT SYSTEMS, VOL. 24, 340-356, 2009.

[6] M. Fernández, I. Cantador , V. López, D. Vallet, P. Castells, E. Motta ," Semantically enhanced Information Retrieval: An ontology-based approach," Web Semantics: Science, Services and Agents on the World Wide Web 9 ,pp. 432-452, 2011.

[7] A. Nawaz and A. Khanum," Ranked Neuro Fuzzy Inference System (RNFIS) for Information Retrieval," Springer, ISNN 2011, Part I, LNCS 6675, pp. 578–586, 2011.

[8] M. A. A. Leite and I. L. M. Ricarte," A Framework for Information Retrieval Based on Fuzzy Relations and Multiple Ontologies," Springer, pp. 292-301, 2008.

[9]J. Zhai, M. Li, and J. Li, "Semantic Information Retrieval Based on RDF and Fuzzy Ontology for University Scientific Research Management," Affective Computing and Intelligent Interaction, AISC 137, pp. 661–668, 2012.

[10] Z. E. Alarab, A. M. Gadallah, H. A. Hefny, "An Enhanced Model For Linguistic-based fuzzy ontology," the 47th Annual Conference on Statistics, computer sciences and operation research, pp. 49-62, 2012.

[11] Y. Bassil, "A Survey on Information Retrieval, Text Categorization, and Web Crawling," Journal of Computer Science & Research (JCSCR) - ISSN 2227-328X , Vol. 1, No. 6, Pages. 1-11, December 2012.

[12] C. D. Manning, P. Raghavan, H. Schütze, "Introduction to Information retrieval: Evaluation in information retrieval," Cambridge University Press, 2008.

Accounting IT systems and requirements of Polish law

Elzbieta Wyslocka and Dorota Jelonek

***Abstract*—** Accounting as one of the oldest economic science has developed over the centuries. With the development of civilization and technological progress also principles, methods and techniques of records used in accounting have changed. This article examines the extent to which computer programs used in accounting, are dependent on the specific requirements of national legal solutions. These requirements are, inter alia, strictly defined in the Polish Accounting Act, but their implementation in the IT system poses many problems. The paper discusses the most commonly encountered.

***Keywords*—** accounting books, IT system,.

I. Introduction

The credibility and reliability of information provided by the accounting systems affects the process of management and control of the organization. It is through the information functions of accounting and the information contained in the financial statements that is possible to perform financial analysis and evaluate achievements that allow to cohesively assess all actions from elementary, individual to particular responsibility centers and the company as a whole. The image quality of a company presented by accounting in its financial statements, is subject to compliance with the principles articulated in the law or professional accounting standards of the country, as well as compliance with the International Accounting Standards.[1, 2]

The Accounting Act currently functioning in Poland approaches universally the accounts kept in electronic form and in the traditional manner. In addition, it attempts to copy the solutions developed in the practice of traditional bookkeeping to electronic bookkeeping. The legislature in the past 20 years has tried to supplement solutions by adapting them to the changing technological capabilities in the field of computerized bookkeeping. Unfortunately, some provisions related to bookkeeping using a computer are somewhat contrary to the copyright laws, such as those concerning the need to specify the algorithms by which a system of a given supplier operates. Therefore, some key questions arise. Should the Accounting Act contain a chapter on bookkeeping? If so - should a number of specific issues be abandoned, and only the minimum requirements defined? If not - how to determine the conditions of bookkeeping and whether determine them at all or leave to the unit discretion?

The article examines the extent to which the computer programs used in the Polish accounting, are dependent on the specific requirements of national legal solutions. Earlier attempts to explain the impact of differences in cultural factors on corporate financial reporting systems have been made. Studies have demonstrated that there are different patterns of international accounting and the development of national systems seems to be a function of environmental factors, which identification raises some controversy as to identify patterns and influential factors [3, 4, 5]. The paper puts the argument that financial and accounting (FA) systems must be subordinated primarily to applicable national laws or international solutions, although the International Financial Reporting Standards do not put too strictly defined requirements in this regard. In contrast, the Accounting Act dated 29 September 1994 (Journal od Law 2013, pos.330) [6] which is in force in Poland and the Position of the Committee of the Accounting Standards Board of 13 April 2010 on certain principles of bookkeeping [7] clearly define the requirements for accounting books, including those conducted in a digital form. These requirements, however, apply to books, without interfering with either method or technology used. Some regulations of the Act are formulated in not fully understandable language; thus it shouldn't be a surprise, that IT specialists have problems with appropriate translation of accountancy principles to algorithms of data processing, although it may negatively affect the quality of the books of accounts serviced by the particular program. Responsibility for adopting the relevant accounting (policy) principles and supervising fulfillment of accounting obligations is held by entity's manager (art. 4 and 4a of the Act). It is beyond any doubt that a choice of appropriate system considerably determines a method of bookkeeping and quality of records, and may significantly influence its clarity, reliability, readability and completeness [8]. Therefore, it is worth at this point to indicate few requirements for FA computer programs, resulting from the provisions of the Act.

II. The Accounting Act and Documentation of Accounting Policy

The obligation to apply the requirements of the Accounting Act relates to entities listed in art. 2 (subject scope of the Act). Indicated unlimited duty applies to most individuals domiciled or operate businesses in the territory of the Republic of Poland, regardless of the amount of revenues and forms of

E. Wyslocka is with the Management Department, Czestochowa University of Technology, Al. Armii Krajowej 19b, 42-200 Czestochowa Poland (phone: +48 601 209 175; fax: 48 34 361 76 38; e-mail: wyslocka@zim.pcz.pl).

D. Jelonek is with the Management Department, Czestochowa University of Technology, Al. Armii Krajowej 19b, 42-200 Czestochowa Poland (e-mail: dorota.m.jelonek@gmail.pl).

business.

The art. 5, paragraph 1 of the Accounting Act requires an entity to select and applied accounting principles (policy) in subsequent financial years in a continuous manner, from period to period in order to ensure comparability of data. This is to ensure uniform grading and, consequently, the assessment of the nature, time of recognition in the accounts and the accounting ledger accounts of particular types of economic events, to apply the same principles of valuation, including the depreciation of intangible assets and fixed assets. In addition, one should ensure a unique grouping of operations or balances in higher sets and showing them in the same positions in the financial statements, as well as the identity of the stock of assets, liabilities and shareholders' equity at the date of closure of the accounts of the preceding financial year to the opening balances of the next year.

Pursuant to the provisions of Article 10 of the Accounting Act, the entity shall have documentation in Polish describing accepted accounting principles (policies), and in particular regarding the determination of the financial year and of its reporting periods, methods of valuation of assets and liabilities and determination of the financial result, the method of bookkeeping, which primarily means:

- company account plan and adopted rules for the classification of events
- a list of the accounting books, and in the case of electronic bookkeeping - a list of data sets representing accounts on computer data carriers with an indication of their structure, interrelationships and their functions in the organization of the whole accounting records and in the processes of data processing
- description of the data processing system, and in the case of electronic bookkeeping - description of the system, including a list of programs, procedures, or functions, depending on the structure of the software, together with a description of algorithms and parameters, and software data protection principles, in particular the methods of securing access to data and processing system, and also to determine the software version and date of commencement of its operation,
- a system for the protection of data and their sets, including accounting documents, books of account and other documents constituting the basis for the entries.

The information system must be fully adjusted to the accounting policies adopted by the user, in particular, to its elements, for which the Accounting Act leaves individuals free to choose the methods of valuation of assets and liabilities and method of determination of the financial result, company chart of accounts, general ledger chart of accounts, adopted rules of events qualification, principles of running subsidiary ledger accounts and their links with the general ledger accounts. Documents related to some of the above indicated rules are included in the accounting records, they are available in the system. Independently of system safety principles, clearly defined rules for the organization of data protection, among other things, describing the company circulation system, numbering and storage of accounting records should be attached to the entity's accounting policy documentation.

Fulfilling the requirements of the Act, an information system provides access to the list of data set representing accounts on computer data carriers with an indication of their structure, interrelationships and their functions in the organization of the whole accounting records and processes of data processing. In addition, the creator of the program should also provide documentation system containing a description of the system, a list of programs, procedures, or functions, together with a description of algorithms and parameters and the data protection principles, including, in particular, methods for securing access to data, and also to determine the software version and date of commencement of its operation.

When bookkeeping is performed with the computer accounting information resources, organized in the form of separate computer data files, databases or its separated parts, regardless of the place of their creation and storage it is considered as equivalent of accounting books. Information resources necessary to maintain the accounting system in the above form is to have software capable of supplying clear information with regard to the entries made in the accounts through their print or transfer to another computer storage media (article 13 paragraph 2 and 3 of the Act).

Separate problems are generated by need to change the accounting system during the financial year. Although the Act does not contain any restrictions on changing FA (financial and accounting) system during the financial year, however it should be remembered, that the FA system should provide information on current balances and transactions of all general ledger accounts and balances subsidiary ledger accounts, thus giving the basis for the existing unit financial statements. The fact that such a change will need to properly test and reconcile data in the revised accounting system and to ensure the merit continuity (comparability) of data provided by the amended FA system and data from previously applied accounting system should be taken into account. The issues that should paid attention when changing accounting system is the chart of accounts and algorithms (calculations) built into the FA system. Changes in algorithms during the financial year, as a violation of the principle of continuity is not acceptable, because at the stage of implementation of the new accounting system it is required to confirm by appropriate tests that the results obtained by the use of modified FA algorithms covered by the new system are consistent with the results of those obtained previously. Documentation of testing algorithms is included in the tests for determining the implementation of the new accounting system; approved by the head of unit. If you change the accounting system during the financial year technical problem is to ensure the presentation of comparable data accrued since the beginning of the year. Therefore, it is imperative that if the accounting system is changed the continuity of data from the beginning of the year until the system changes is ensured. Data from the current accounting system can be transferred to a new accounting system through: detailed records, accounts, or balances. The transfer of all records secures detailed presentation of data on a continuous

basis, but it is rarely used (in particular, if the accounting system change involves changing the chart of accounts, the transfer of records details is usually not possible). In general, the transfer of data from the period preceding the change to the new accounting system is done on an aggregate basis, i.e. turnover or balance. When transferring aggregated data the key problem is to document relations between the detailed records (in the current system) with the aggregated records (in the new system).

Changing the existing accounting system to a new, requires to ensure that the new system allows adequate protection of both the data and the software and the planned structure of the permissions on the new system is suitable. Ancillary data used during the transfer of data are archived on a durable medium and protected against loss (archived) on general principles.

III. Electronic Document as an Accounting Evidence

Accounting entry is a fundamental component of the books of accounts. As a result of processing data covered by a book entry, one can obtain specifications presenting numerical data in chronological order (journal) or systematic order (accounts). In order to obtain specifications presenting appropriately data recorded in the books it is necessary to ensure:

- completeness of information representing single accounting entry,
- creation of accounting specifications exclusively on the basis of verified and recorded accounting documents,
- an effective way of linking accounting entries with underlying source of documents.

The term "electronic document" refers to the form of a document, such as: electronic invoice, print from the online banking system, electronic storage document, electronic ticket data from billing systems, etc. An electronic document representing an accounting evidence should comply with the requirements set out in article 21, paragraph 1 of the Act for the accounting evidence, including the simplifications referred to in art. 21, paragraph 1a of the Act. When bookkeeping using the computer the accounting records, made automatically through communication devices, computer storage media or generated by an algorithm (program) on the basis of information already in the books are considered equivalent of the source evidence (Article 20 paragraph. 5 the Act). Such provisions may also occur as a result of the introduction to the books of electronic documents constituting evidence of accounting. Entries entered automatically into the books of accounts (accounting system) shall be considered equivalent to provisions made on the basis of the source evidence, if they meet at least the following conditions:

a) when registering they become permanently readable and compatible with the contents of the relevant accounting documents;
b) it is possible to determine the source of the records and the person responsible for their introduction to the books of accounts and further modification;
c) the applied procedure provides validation of processing of relevant data and the completeness and identity of records;
d) source data in place of their creation are properly protected in a way that ensures their persistence, for the period required to store a given type of accounting documents.

Accounting entries are made by a computer system in a sustainable manner, without leaving places to use later insertions or changes. The system provides protection for records against their destruction, modification or covering of entry. In addition, the records in the log and ledger accounts are linked in a way that allows checking their compatibility. Information system provides storage of records in the accounts for a period of not less than required in the Accounting Act.

Each accounting document is defined in the computer system by: type of evidence, identification number, the parties engaged in a business transaction, a description of the operation, date of operation and preparation of evidence.

These signs provide identification of accounting evidence and comply with the requirements of the Accounting Act. Accounting documents generated by the online system meet the conditions under the Accounting Act and the VAT Act for documenting events subject to VAT.

The collected documents, regardless of the place of origin, may be transferred, edited and accepted within the defined paths of the electronic circuit. Existing "migration" of paper information rules may be replaced by their electronic version, assuming their compliance with the original. It accelerates the way of operations of the company and improves the efficiency of its activities. As a result of the transition to electronic workflow employees perform tasks in a shorter period of time, the team effectively uses the gathered jointly information and documents, and the company reduces costs resulting from normal transmission and archiving of documents. Programs for reading paper invoices that analyze scanned documents and automatically add to the information on the invoice are more and more commonly used. This allows you to export them to a file, import and proper entry in the accounting program, which is used by the company. An archive is automatically created and allows for the resignation of bulky paper files and searching for documents with the help of searcher and browse them directly in the accounting program. These applications are usually available on-line, so it does not require any additional installation, investment in hardware or expensive servers. Conditions to use them are an Internet access and a web browser.

Criteria for evaluation of completeness of information, which are a basis for single entry, are determined by art. 23 paragraph 2 of the Act. Programs, which accept possibility of delayed entry should contain also one information – the date of actual accounting entry, filled automatically with calendar date according to system clock.

IV. Accounting Specification and Accounting Act

The Act clearly defines the criteria for properly designed accounting books, which are all subject to the accounting records. Facing requirements of the law IT system allows for

keeping the books in a continuous manner, and enables making the accounting records up to date. The functionality of the system provides access to data sets which can be granted at any time and any place.

The user can use various functions of accounts such as Dr/Cr turnover, Dr/Cr cumulative turnover, per balance, opening balance opening balance difference**,** and others. Extremely important is the ability to create any combinations based on the chart of accounts, high level of integration with Microsoft Excel provided by most FA programs allows to use tools such as pivot tables and charts**,** raw data, charts, and tables loaded to the system, there is no need to create an update of prepared statements because they are subject to dynamic updating of the database system, ideal for those with only a basic knowledge of MS Excel**,** as well as "Excel" professionals, who can not imagine working without these tools.

Art. 14 of the Act determines the content and organization of the journal. In order to present accounting entries in chronological way, algorithms of financial/bookkeeping program should contain a fixed expression of chronology. A choice of appropriate method seems to be very difficult for many designers of financial/ bookkeeping programs. The authors of software use various solutions for assuring conformity of the date of entry in the books with automatically prescribed number of journal. Some sample solutions are provided below:

- Positions in the journal are determined at the first registration of data which facilitates noting in the document a number of position in the journal but does not guarantee full consistency of numbering with the date of entry predetermined by the operator; it does not protect against gaps in numbering in case of removing entry from the system's resources as well.
- During the first registration, the program assigns a subsequent number of a journal to *documents* allowing multiple changes of the number with regard to a date considered in a program as an indicator of chronology up to the closing of reporting period; all gaps in the numbering resulting from the canceling of wrongly registered positions are then eliminated. In this situation, continuity of numbering is preserved, but it is impossible to use a number of a journal position as an identifier which matches IT resources with an archive of *source documents*.
- When the *document* is registered for the first time, the program sets a subsequent record number of accounting document (the number is blocked from editing), but only after a final acceptance of a record, the proper position number is assigned in the journal. Record number can be noted on the *document* which helps to link it with appropriate *source document*; it also guarantees a constant numbering of positions in the journal.

A structure and frequency of trial balances of general ledger accounts is determined by art. 18 paragraph 1 of the Act. The trial balance should obligatorily include a complete list of used parameters, determined before the specification is prepared. If the list is missing, the work can be indeed much more difficult. The trial balance for a particular reporting period should include information whether the books for that period has been finally closed. Otherwise, one cannot be sure, if balance values in specification are definite or present only a temporary situation.

Entering data into the accounting system and deciding how to classify these data in the accounts are two different things. Activity concerning entering data into the accounting system is associated with saving data in a given way. In contrast, decisions about how to classify the data to include in the accounts fall within the scope of control over the accounting books and records in them. In particular, some computer financial and accounting systems allow to separate the introduction of source evidence to the accounting system since from including it in the accounts. If the introduction of evidence of accounting data to the accounting system is not directly related to the:

a) deciding on the way of classifying it in the accounts;

b) control of the accuracy of events recording, made in the accounts - the input of data to the system itself is not equivalent to keeping books of accounts.

Keeping records in the edit state (e.g., before accepting the buffer) can not be regarded as identical with the recognition in the accounts qualified for entry in the accounts in the month of accounting documents

In article 12 of the Act certain requirements on the closure of the accounts for the financial year have been presented. Closing of the reporting period in the computerized accounting usually means changing the status of the period from active (open / active) to the closed (locked). The power to do this are usually assigned to users with the highest rank - system administrators. The closure of period should block any possibility of introducing new records to the closed books. Some programs distinguish between temporary and definitive closure of period. The temporary closure is applied to the time of verification and reconciliation of accounts for the closed period. Closing the final (irreversible) should take place immediately after filing (and possibly auditing) of all reports for that period. Books for the whole year should be locked, preventing any changes to the cycle defined in the given paragraph 4 of art. 12 of the Act. The lack of such functions may undermine the accuracy of the books.

V. Accounting in the Cloud as a Bookkeeping

Accounting Act in Article 11 allows for the possibility of entrusting bookkeeping to the entrepreneur conducting business in this field as well as within the entity using its information, organizational and human resources. As already mentioned, in accordance with Article 13, paragraph 2 and 3 bookkeeping using the computer appropriate accounting information resources, organized in the form of separate computer data files, databases or separated parts of it, regardless of the place of their creation and storage are considered equivalent of accounting books.

The books of account carried out using a server that is outside the place of keeping the books, is considered to be carried out in a proper manner, if they meet at least the following conditions:

a) entity has control over the accounting books and records made in them;
b) the entity secures identity of the accounting records with a copy of the reports received by teletransmission connection (wired) and wireless;
c) the accounts are conducted fairly, correctly, verifiable and up to date,
d) a clear link between the accountancy records with supporting documentation is provided;
e) the books of account are effectively protected against unauthorized changes, unauthorized access, damage or destruction;
f) the books are available at all times at the place of bookkeeping.

In a situation where the system stores the vouchers only in the electronic version, they can be accessed using servers located in the unit or outside the unit, which will be owned or leased by the entity. "In the cloud" data processing came to be called a model of the IT systems in which the server installation location does not matter. Cloud computing is a new model of computation that can bring significant benefits to consumers, businesses and government, creating new threats and challenges. Accounting "in the cloud" is a relatively new phenomenon. Accounting is a field rather conservative and one of the last subjected to modern computer and technological trends.

A few years ago the company, at the expense of considerable amounts, invested (and still invest) in data centers or server rooms, equipped with a sufficient number of servers that perform different functions. To support server administrators security companies have been employed and engaged. With the introduction of new solutions in recent years, it turned out that the concept of building their own data center is not always effective. An important element is to increase awareness among entrepreneurs advantages of outsourcing, or outsourcing certain services to external partners. Accounting will also be aimed in the direction of the "in the cloud" data. The factors reinforcing this trend will be publicly available service on the Internet and the integration of accounting systems with these services.

VI. Data protection, Including the Protection of Personal Data

The issue of retention of documents is regulated by section 8 of the Act of accounting "Privacy protection". In accordance with art. 71 books of accounts, accounting documents, including e-invoices, inventory, and financial statements must be properly stored, protected from prohibited against amendments, distribution, damage or destruction. The range of issues related to information security include:

- procedures for granting authorization to process data and registration of these rights in the information system;
- identifying the person responsible for the procedure of granting authorization;
- the means of authentication and procedures related to their management and use;
- procedures for starting, suspension and termination of work by the users of the system;
- procedures for backup data sets and programs and software tools used for processing;
- manner, place and period of storage - electronic media containing personal data backup.

All collections of documents generated by the system (accounting books, accounting documents) are stored and secured on servers from intrusion. Servers provide 24-hour access to accounting data unit and minimize the risk of data loss due to a malfunction. In addition, every day creates backup copies of data files that are stored on computer media and stored for a period specified in the Act (5 years, excluding financial statements which should be stored permanently).

Most software allows to define certain parameters of setting the password such as: minimum password length, password complexity checking rules, forcing the user to change the password after a certain date, etc. To ensure the highest level of data protection to observe basic security policy requirements.

The information system provides access to accounting data unit only to third parties, the manager of the unit expressed written authorization of access to the system. As the online system provides access to accounting data of company for many previously identified users, automatic monitoring system keeps records made by individual users. Every document created in the system is marked by signature of the user who entered it. Each authorized user has separate roles (accountant assistant, accountant, independent accountant, chief accountant), the right of access to data (total absence, read-only, full access to the data), the system of passwords is closely associated with the security policy in your company.

The Regulation of the Minister of Internal Affairs and Administration of 29 April 2004 on personal data processing documentation and technical and organizational conditions which should be fulfilled by devices and systems used for the processing of personal data [9] imposes the obligation of the data controller to develop instruction specifying the management of the computer system used to process personal data, its approval and transfer of procedures and guidance contained in the manual to those responsible for their implementation.

VII. Conclusion

Financial and accounting programs must be subordinated primarily to national laws or international solutions being in force. First of all functioning in the enterprise information system must be fully adjusted to the accounting policies adopted by the user. This is especially true for those solutions for which the user itself makes a choice in accordance with the laws of the balance sheet. The necessary condition for obtaining the statements which in the correct way describe the data contained in the books is to provide the completeness of the information constituting the single entry, preparation of accounting statements only on the basis of the verified and recorded evidence, and to provide an effective way to link

accounting records with source documents. In the era of globalization and performance of transnational enterprises, availability of current financial information from anywhere in the world and at any time, becomes a necessity. That data processing concerning costs, revenues, sales, corporate finance in the cloud enables independent of place and time access to such data only limited by access privileges. However, new information solutions cause new legal problems in terms of the operation of information systems in the field of accounting in relation to the increasingly common use of Cloud computing. The basic condition for exploiting benefits of cloud computing activities, not only for the single market of the European Union, is to fill the gaps in the legal provisions related to the cloud. Main points are to improve conditions for users, solving security-related problems of stakeholders, to encourage the public sector to benefit from the cloud and to support further research and development in cloud computing

REFERENCES

[1] A. Jaruga, "Harmonizacja rachunkowości u progu XXI wieku," in *Rachunkowość u progu XXI wieku*, L. Kopczyńska, Ed. Warszawa: Fundacja Rozwoju Rachunkowości w Polsce, 1998, p. 62.

[2] E. Wyslocka, *Informative Function of Accounting and Information Technology Systems*, Electronic Modeling, vol. 29, no. 4, pp. 129-137, 2007.

[3] S. J. Gray, *Towards a Theory of Cultural Influence on the Development of Accounting Systems Internationally*, Abacus, Vol. 24, Iss. 1, pp. 1–15, March 1988.

[4] R. D. Nair and W.G. Frank, *The Impact of Disclosure and Measurement Practices on International Accounting Classifications*, The Accounting Review, July 1980.

[5] C.W. Nobes, *A Judgemental International Classification of Financial Reporting Practices*, Journal of Business Finance and Accounting, Spring 1983.

[6] The Accounting Act dated 29 September 1994 (Journal of Law 2013, pos.330).

[7] The Position of the Committee of the Accounting Standards Board of 13 April 2010 on certain principles of bookkeeping, resolution no. 5/10.

[8] E. Wyslocka, *IT Systems Supporting Finance and Accounting*, Electronic Modeling, vol. 31, no. 6, pp.87-98, 2009.

[9] The Regulation of the Minister of Internal Affairs and Administration of 29 April 2004 on personal data processing documentation and technical and organizational conditions which should be fulfilled by devices and systems used for the processing of personal data (Journal of Law 2004, no.100, pos. 1024).

Integration of open source systems for visibility of scientific production of universities

Ionela Birsan, Daniela Drugus, Marius Stoianovici, Angela Repanovici

***Abstract*—** The article shows us the research conducted in order to find the barriers to dissemination and communication of scientific articles published by such institutional digital repositories. Transilvania University of Brasov, Romania, provides a new service attached to the digital repository, an automatic query interface of the SHERPA RoMEO platform, a publisher reviewing platform. Pressing the new button will open a new window (pop-up) of the browser, where the actual query of the SHERPA RoMEO server will be performed then, upon further closing the window, the user will automatically switch to the next page – the second step of submitting the new item (where account is taken of the options being checked in the first page, even if the newly created button was pressed instead of the "*Next*" button). The transformation results in a HTML file. It is a simple list of the identified publishers and the "*romeo*" color associated with each of them. The software application developed with very low cost price can be also used as a model for other universities. The application is original, the model is easy to develop, and the practical implications are of great use to the academic community**.**

***Keywords*—** Digital repository, DSpace, Romania, SHERPA RoMEO, XML, XSL.

I. Introduction

THE movement of open access to information has developed new models of communication and dissemination of scientific information. Universities have provided the academic community with instruments to promote scientific production, creating institutional digital repositories. The academic community reacted with disbelief, the number of archived documents being below expectations. Most authors are afraid of breaching copyright and do not want to devote time to reviewing the publishing conditions imposed by publishers.

Many surveys have been conducted on the academic community's behavior concerning the Open Access movement and the motivations and impediments for which they archive or not their papers published in open access.

Ionela Barsan is with Transilvania University of Brasov, Romania (e-mail: one.barsan@yahoo.com.)

Daniela Drugus is with University of Medicine, Iasi, Romania(e-mail: daniela.drugus@umfiasi.ro).

Marius Stoianovici is with Transilvania University of Brasov, Romania(e-mail: marius_stoianovici@yahoo.com)

Angela Repanovici is with the Product Design and Environment Department at Transilvania University of Brasov,Romania(e-mail: arepanovici@unitbv.ro, telefon 40745820361, corresponding author).

The vast majority of surveys refer to the behavior of researchers on archiving in institutional digital repositories. [1], [2].

The main barriers encountered are: copyright concerns [1], [3], [4], [5], [6], additional time and effort [7] (Van House, 2003), mistrust [8].

Faculty with technical skills and younger faculty are more involved in self-archiving articles. Providing logistical and technical support will also foster participation of those who are less computer adepts [9],[10],[11],[12].

Issues relating to copyright and intellectual property are also generated by not knowing the conditions within the publishing agreements. There is misconception that self-archiving breaches copyright agreements [13].

Authors fear that, by the publishing agreements entered into, they are not allowed to upload their papers in institutional digital repositories. One in ten authors knows such problems, while the other nine have only "a slight idea". This ignorance leads to a tendency for authors to be over-cautious [14].

Copyright remains the biggest obstacle in self-archiving articles in institutional digital repositories [15], [16], [17].

II. Implementing and Using the DSpace Platform

DSpace [18] is a software application (platform) designed for academic, non-profit and also commercial organizations with a view to developing and managing digital repositories. The digital repository is a collection of digital documents, organized in a well-defined hierarchical structure. DSpace software is free of charge and easy to install (*out of the box*), and fully customizable, in order to suit any organization's needs.

DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images (video) and data sets. With a growing community of developers committed to continue the software's expansion and improvement, each application installation benefits from the experience of the previous users and developers.

DSpace is the software support of a digital repository of documents. In turn, the digital repository is the environment (software) where the *institutional digital repository* may be created. There are several definitions of the "institutional repository". Lynch [19] defines the institutional repository as: "a set of services that a university provides to its community members, for the management and dissemination of digital materials created by the institution or by the members of such

ISBN: 978-1-61804-236-1

community. Organizational commitment is essential to manage such digital materials, including long-term preservation where appropriate, as well as organization, access or distribution". Ware [20] also includes the participation of the open archives initiative (OAI - Open Archives Initiative [21]): "a web-based database (repository) of scholarly material which is institutionally defined (as opposed to a subject-based repository); cumulative and perpetual (a collection of record); open and interoperable (using OAI-compliant software); collecting, disseminating and storing (is part of the process of scholarly communication). In addition, most would include long-term preservation of digital materials as a key function of the institutional repository".

For DSpace to become an institutional repository, special attention should be given to configuring and managing the same. But DSpace can be the support of any other type of digital repository, the "institutional" or "subject-focused" or any other character of the repository being brought about by the way the software application is further configured and managed.

Almost 1,500 installs of the DSpace software platform are currently known worldwide, most of them in the academic environment [22]. Out of these, 9 are in Romania, of which 6 are in the academic environment (one of them being the *Aspeckt* platform of Transilvania University) [22],[23]

III. SHERPA RoMEO PROGRAMMING INTERFACE

RoMEO is a database of publisher copyright policies on self-archiving, based on the publisher's copyright transfer agreement. It is maintained by SHERPA with support by JISC and the Wellcome Trust. Individual journal titles, ISSNs or publishers can be searched, and each title is identified as Green (can archive pre-print and post-print), Blue (can archive post-print (i.e. final draft post-refereeing), Yellow (can archive pre-print (i.e. pre-refereeing), or White (archiving not formally supported).

API (*Application Programming Interface*) is an acronym used to generally refer to collections of predefined software functions that allow writing custom applications running in a predefined environment. The application programming interface for SHERPA RoMEO is a machine/machine interface allowing programmers' access to Sherpa Romeo data within their own developed applications. For example, API can be used to embed automated searches of journals or publishing houses during a record (submission) process of a paper in a repository.

Like most APIs used in the web environment, the SHERPA RoMEO interface does not involve downloading a library of functions on the user's computer, but calling functions from a web application server, by HTTP queries.

IV. INTEGRATING SHERPA/ROMEO WITH DSPACE

Integrating information contained in the Sherpa/Romeo database into DSpace platform can be made by combining all the information presented in the previous chapters. A general diagram of the process is shown in Fig. 1.

There are two versions of the XSL language used in practice:

- XSLT [24], defining the transformations being applied to the XML tree;
- XSL-FO (*Formatting Objects*), used to transform XML documents into binary format documents such as PDF or even Microsoft Word.

There are three ways that an XML document may be transformed into another type of document by applying an XSLT stylesheet:

- the XML document and the associated stylesheet are sent to the client application (browser) whose task is to effectively perform the transformation according to the information in the XSLT stylesheet. In such conditions, server load decreases but the browser should allow processing of XML documents;

- applying the XSLT stylesheet is carried out on the server itself, the resulting document (usually in HTML format) being sent to the client. Thus, processing may be carried out according to the nature of the client program;

- the third possibility is very rarely used and refers to the transformation of the XML document using an external application and placing the resulting document (HTML) on the server, being further sent to the client.

The core element of the XSLT technology is the template: <*xsl:template*>. Two important elements may be found herein: the *match* attribute – specifies a path to the input tree; the content – implements the way transformation is performed.

The general form of a template is:

```
<xsl:template match="element_XPath">

...

<xsl:template>
```

The association of an XML document with an XSLT stylesheet is performed within the XML document by the processing instruction <*?xml-stylesheet*>*:*

```
<?xml-stylesheet href="stylesheet/Login" type="text/xsl" />
```

The *href* argument specifies the name of the XSLT stylesheet and, where appropriate, the path thereto.

Identifying the fields within the tree structure of the XML document is performed through *XPath* elements (which is sometimes described as a language, although it is not a language proper). The XPath convention is similar in functionality to navigation through the directory structure in the operating systems, such as MsDOS, Linux or Windows.

At conceptual level, at the basis of the XML document's structure (but having no corresponding element within XML elements), is the root of the document, represented by the "/" character.

XPath expressions are interpreted from left to right, for example, for an XML tag which is at the first level of the tree structure (for example <*basis*>), the expression that reads the

element value is "*/basis*", and for the following levels, it could be, for example "*/ basis/level1/level2*" and so on. The previous expression can be understood as: "starting from the document root, select the <*basis*> element, which is its child (the root's)." Failing to write the "/" character in the previous XPath expression radically changes the meaning of such expression, in which case "all the <*basis*> elements, which are children of the current node are selected". In the case of more complex XPath elements, the constituents are separated by the "/" character, which, as can be noticed, has a double meaning depending on the position in which it appears within the XPath element.

Moreover, there are certain situations where an XPath element must make a much more rigorous selection of the elements selected and treated within a template. It can be assumed, for example, that in a certain context only selecting the <input> elements is wanted, whose "type" attribute has a value different from "hidden". In order to achieve this, the element which will be filtered must be followed by the filter to be applied. It consists of a pair of brackets ([]), which usually frames a condition.

While defining a new submission for a collection on the DSpace platform, a query is sent to the Sherpa/Romeo server, using the application programming interface it provides, which responds by an XML document containing the information required. The XML document is then processed through the XSLT transformation so as to generate the content displayed on the HTML page.

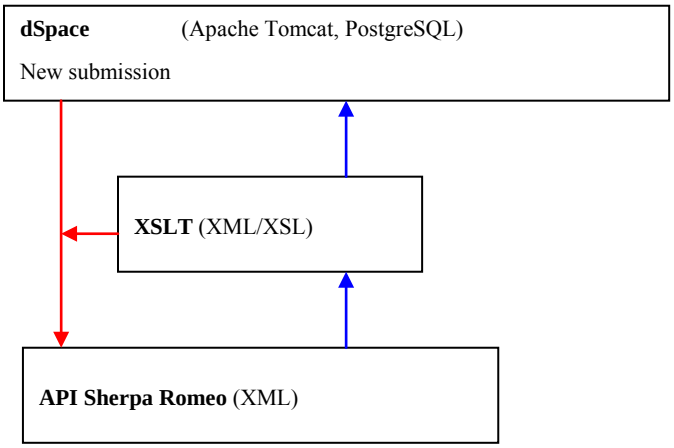


Fig. 1 Diagram of integrating Sherpa Romeo with DSpace

These actions involve taking control from the original application of the DSpace platform and inserting a page in the chain of submitting a new item on the platform. As the DSpace platform is developed by using the JSP (Java Server Pages) technology, namely the HTML pages are generated using functions written in Java language, the sequences newly introduced in the record chain should be preferably written in the same language, using the same JSP technology.

V. Submitting the Request

The first step in recording a new article consists of the interactive setting of several variables influencing the way in which the following pages are displayed: number of versions of the article title, the number of files to be uploaded and whether the article was published before (in this case, specifying the publisher will be requested).

In this first page, intervention may be made on the JSP code to create a new button, as shown in Fig. 2. The newly-created button is "Check Publisher", circled in red in the figure.

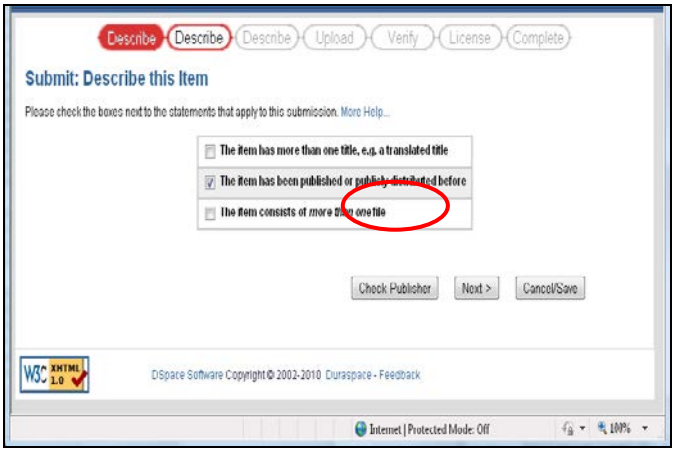


Fig. 2 Diagram of integrating Sherpa Romeo with DSpace

Where opening the window is triggered on the onclick action (mouse click on the button) (DHTML), and the transition to the next page is made by the input tag of submit type.

The page opened upon pressing the button will include the query options: ISSN, Publisher, Journal Title.

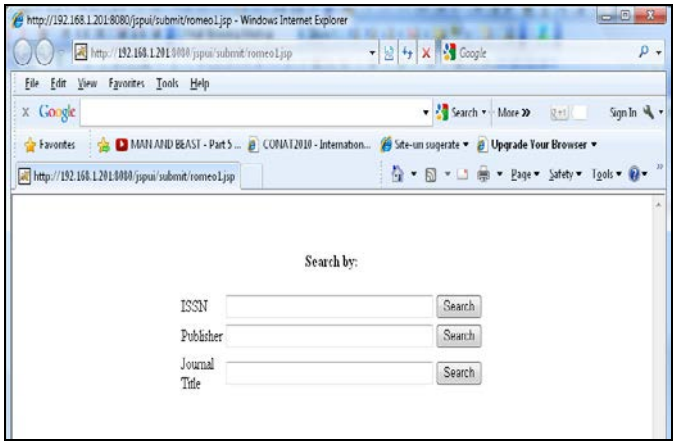


Fig. 3 Simplified search page

A simple search page by the three criteria is the one presented in Fig. 3. It is basically a form with three input fields and three different submit-type buttons, one for each search criterion. Pressing each button sends the corresponding query to the SHERPA RoMEO server, the parameter being the one specified in the input field concerned.

The transformation results in an HTML file that will look like in Fig. 4. It is a simple list of identified publishers and the "romeo" color associated with each.

With a view to obtaining all the information about the publisher, it should be searched either the exact name of the

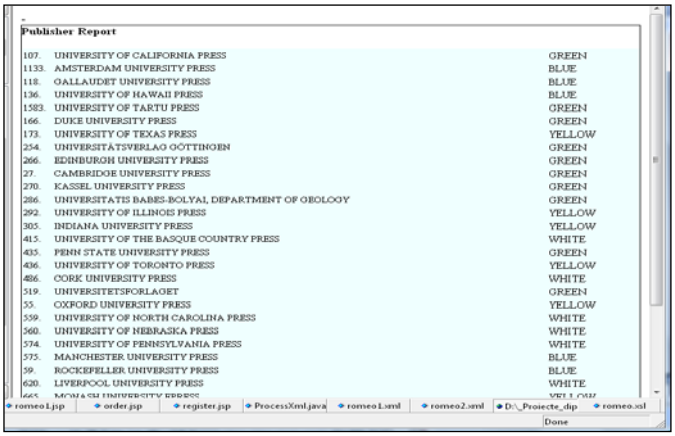


Fig. 4 Search result by the publisher name

publisher, but it is rather likely to be introduced differently from the Sherpa submission, or the ISSN code. The search by title equally displays a list, as a result, as if searching by the publisher name. In both cases the processing may be extended by the user selecting an item from the list.[25]

VI. Conclusions

The software application has practical implications and represents an original solution to the needs of the academic community. The application is created with a very low cost, both platforms are free to use, being the results of research projects. The proposed model can be also very useful to other universities with the same problems and obstacles in populating and developing institutional digital repositories. A practical implication concerns the easy, one-button, access to two applications simultaneously: archiving, in a digital platform, a published article and accessing the list of publishers enrolled in the platform SHERPA RoMEO's database of publisher policies on open sharing. Another implication is to reduce the time of promoting the digital repository's services and the archiving time. Regarding originality and article value, it is ascertained that a need identified in the self-archiving process is solved, a barrier to the use of digital repository through an original software application.

ACKNOWLEDGEMENT

This paper is supported by the Sectoral Operational Program Human Resources Development (SOP HRD), ID134378 financed from the European Social Fund and by the Romanian Government.

References

[1] J. Allen, "*Interdisciplinary differences in attitudes towards deposit in institutional repositories* (unpublished master's thesis)", Manchester Metropolitan University, UK, 2005.

[2] S. Watson, "Authors' attitudes to, and awareness and use of a university institutional repository". In *Serials*, 20(3), pp. 225-230, November 2007.

[3] L. Chan, "Supporting and enhancing scholarship in the digital age: The role of open-access institutional repositories", in *Canadian Journal of Communication*, 29(3), pp. 277–300, 2004.

[4] E. Gadd, C. Oppenheim and S. Probets, "RoMEO studies 1: The impact of copyright ownership on academic author self-archiving", in *Journal of Documentation*, 59(3), pp. 243–277, 2003.

[5] E. Gadd, C. Oppenheim and S. Probets, "RoMEO studies 2: How academics want to protect their open-access research papers", in *Journal of Information Science*, 29(3), pp. 333–356, 2003.

[6] N. F. Foster and S. Gibbons, "Understanding faculty to improve content recruitment for institutional repositories", in D-Lib Magazine, 11(1), 2005.

[7] N. A. Van House, "Digital libraries and practice of trust: Networked biodiversity information", in *Social Epistemology*, 16(1), 99–114, 2002.

[8] R. Crow, "The case for institutional repositories": A SPARC position paper, in *ARL Bimonthly Report*, 223. Retrieved July 6, 2010, from http://works.bepress.com/cgi/viewcontent.cgi?article=1006&context=ir_research

[9] J. Kim, "Faculty Self-Archiving: Motivations and Barriers", in *Journal of the American Society for Information Science*, 61(9), 1909-1922, 2010.

[10] A. Comsa,I. Maniu,, N. Modler, W. Hufenbach, W., EC Lovasz,, V. Ciupe, An Overview of Library Automation in *Mechanisms, mechanical transmissions and robotics Book Series: Applied Mechanics and Materials,* (162) 583-588 , 2012

[11] E.C.Lovasz, D. Perju, KH Modler, A.E. Lovasz, I. Maniu, C. Gruescu, D.Margineanu,V. Ciupe, ,A. Comsa, Demonstrative Digital Mechanisms Library in Mechanisms, *Mechanical transmissions and robotics Book Series: Applied Mechanics and Materials* (162), 37-46, 2012

[12] V. Ciupe, EC Lovasz, CM Gruescu, High Quality Document Digitization Equipment, in *Mechanisms, mechanical transmissions and robotics Book Series: Applied Mechanics and Materials* (162), 589-596 , 2012

[13] S. Harnad, "Maximizing Research Impact Through Institutional and National Open-Access Self-Archiving Mandates", in *Proceedings of CRIS2006. Current Research Information Systems: Open Access Institutional Repositories*. Bergen, Norway. Jeffrey, K., Eds., 2006 http://eprints.ecs.soton.ac.uk/12093/

[14] A. Sale, "The acquisition of open access research articles", in *First Monday*, Vol. 11 No. 9, October 2006.

[15] F. Singeh,, A. Abrizah and A. Karim, "Malaysian authors' acceptance to self-archive in institutional repositories", in *The Electronic Library*, Vol. 31 No. 2, 2013, pp. 188-207, 2013.

[16] A. Abrizah, "The cautious faculty: their awareness and attitudes towards institutional repositories", in *Malaysian Journal of Library and Information Science*, Vol. 14 No. 2, pp. 17-37, 2009.

[17] K. V. Stanton and C.L. Liew, "Open access theses in institutional repositories: an exploratory study of the perceptions of doctoral students", in *Information Research*, **17**(1) paper 507, 2012. [Available at http://InformationR.net/ir/17-1/paper507.html]

[18] DSpace, Project webpage, 2013 www.dspace.org [accessed on June 3rd, 2013].

[19] C. A. Lynch, Institutional repositories: essential infrastructure for scholarship in the digital age. portal: Libraries and the Academy, 3.2: 327-336, 2003.

[20] M. Ware, "Institutional repositories and scholarly publishing", in *Learned publishing*, 17(2), pp. 115-124, 2004.

[21] P. Dietz, *DSpace 1.7.1 System Documentation. Portal*, www.dspace.org,

[22] ASPECKT DSpace, Transilvania University Institutional Repository Web Page, 2013 http://aspeckt.unitbv.ro, [accessed on June 3rd, 2013]. available at: http://eprints.utas.edu.au/388/1/FirstMondayOct06.pdf

[23] L. Kristick, "Using journal citation reports and SHERPA RoMEO to facilitate conversations on institutional repositories", in *Collection Management*, 34(1), pp. 49-52, 2009.

[24] A. Repanovici, "Intelectual property and open access to information", in *International Conference on intellectual property and information management* (IPM '11), A. Repanovici, and C. Murzea Eds. Transilvania University of Brasov, Romania, 7th – 9th April 2011, pp. 153-158.

[25] Sherpa ROMEO, Project webpage, (2013) http://www.sherpa.ac.uk/romeo/api.html , [accessed on June 3rd, 2013]

Remote Access to RTAI-Lab Using SOAP

Zoltán Janík and Katarína Žáková

***Abstract*—**The paper presents a new unified interface for remote access to the functionality of systems based on RTAI (Real-Time Application Interface). The described solution enables users to design custom schemes locally and compile those remotely using provided web services. Thus, the new functionality creates an environment that allows not only the remote execution of pre-defined real-time tasks, but it offers the ability to create custom tasks remotely as well. Simple integration into existing web applications is ensured by using WSDL (Web Services Description Language) and SOAP (Simple Object Access Protocol) technologies.

***Keywords*—**computer aided engineering, control education, on-line control, real-time tasks.

I. Introduction

REMOTE and virtual experiments in education process are constantly gaining higher importance, therefore many universities put great effort in development of remote and virtual laboratories. Institute of Automotive Mechatronics at Faculty of Electrical Engineering and Information Technology, Slovak University of Technology is not an exception. The most common solution is on-line access to the systems that are accessible for students during standard daily courses. In this way, each system acquires the added value of system availability besides the time that is reserved for teaching. By adding more features to web-based laboratories, on-line experiments gain higher level of interactivity and extended competitiveness to standard hands-on experiments of daily courses.

We have focused on development of unified solution that extends standard on-line laboratories to be able to provide real-time control and remote access by using PaaS (Platform as a Service) model. The developed solution offers a package of modules that can be integrated to systems that realize remote and virtual experiments and provide on-line access to laboratory infrastructure. The most important module is used for remote access to experiments using hard real-time control with enhanced interactivity that exceeds the available alternative solutions.

On-line laboratories that contain tools for real-time based experiments (e.g. Simulink Coder, also known as Real-Time Workshop in Matlab, Scicos-HIL and other alternatives) depend on host operating system. Thus, in addition to the real-time support in applications, the real-time support in operating system is also necessary. General purpose operating systems such as Windows, Mac OS or standard Linux distributions do not match this requirement [8]. For this reason, it is necessary to use operating system with special modified kernel. One of the possible choices among open-source solutions is Linux-RTAI (Real-Time Application Interface) kernel that supports task execution in hard real-time mode. The RTAI kernel is a part of the Real-Time Suite that will be discussed in next section.

Our solution covers a gap in current state-of-the-art of the Real-Time Suite. Whereas there is a way to access the real-time task remotely using RTAI-XML server extension, the remote real-time task has to be provided in advance by the system administrator. This task was fully automated and it was delegated to the web application that runs within the real-time server powered by RTAI platform. Referring to available sources, such solution has not been deployed on any similar project of remote laboratory.

II. Real-Time Environment

Standard installation of Real-Time Suite contains a patch for Ubuntu's kernel that adds hard real-time capabilities to the system. The RT Suite also offers:

- Comedi drivers for real-time communication with hardware (e.g. data acquisition boards),
- ScicosLab and Scicos for design of block schemes (similar to Matlab and Simulink),
- RTAI-Lab and Comedi-lib toolboxes for ScicosLab that allow use of specialized real-time blocks in Scicos,
- RTAI-XML server that makes compiled real-time task available via Internet using XML-RPC (Extensible Markup Language – Remote Procedure Call) protocol.

We have developed a web-based client application that is able to communicate with such real-time tasks [9].

Despite the ability of RTAI-XML server to provide Internet access to real-time tasks, the only way to create custom tasks was to design block schemes in Scicos, compile them locally on the Real-Time server and set-up the RTAI-XML configuration to allow remote connection to the newly compiled real-time task.

It was necessary to extend the standard functionality of Real-Time server powered by Real-Time Suite to be able to compile block schemes remotely using RTAI-Lab toolbox for

This work was supported by the grant "Program for support of young researchers" of Slovak University of Technology. It has also been supported by the Slovak Grant Agency, Grant KEGA No. 032STU-4/2013 and APVV-0343-12.

Z. Janík and K. Žáková are with the Institute of Automotive Mechatronics of Faculty of Electrical Engineering and Information Technology, Slovak University of Technology, Bratislava, Slovakia (e-mail: zoltan.janik @ stuba.sk, katarina.zakova @ stuba.sk).

ScicosLab. We have developed a web service interface to communicate with client applications using SOAP (Simple Object Access Protocol). The architecture of the described environment can be seen in Fig. 1.

The standard installation of Real-Time server had to be extended by Apache Web Server and PHP interpreter to be able to provide web services interface. Also, the MySQL database server had to be installed to provide storage for user and scheme settings and logs.

III. Web Services Provider

Our solution of the Real-Time server contains a component called WS Provider. This component is essentially a web application that communicates with clients using SOAP protocol. The WSDL (Web Services Description Language) file that accurately describes all services is also available. Thanks to the WSDL, the user can use a generic (general purpose) SOAP client, thus there is no need for difficult development of client applications. However, we recommend using the developed module that is dedicated to provide communication interface with WS Provider (see section IV).

The WS Provider component enables clients to submit an XML file containing the custom block scheme. Compatible XML files can be exported directly from desktop version of Scicos (using any operating system) or from provided web-based scheme editor module. The submitted scheme is processed on the Real-Time server afterwards. Each scheme is checked for presence of potentially malicious blocks or code that could harm the server or hardware equipment. After a successful verification, the submitted scheme is inserted into the framework scheme. The selection of the desired framework scheme is done by the user before the custom scheme submission.

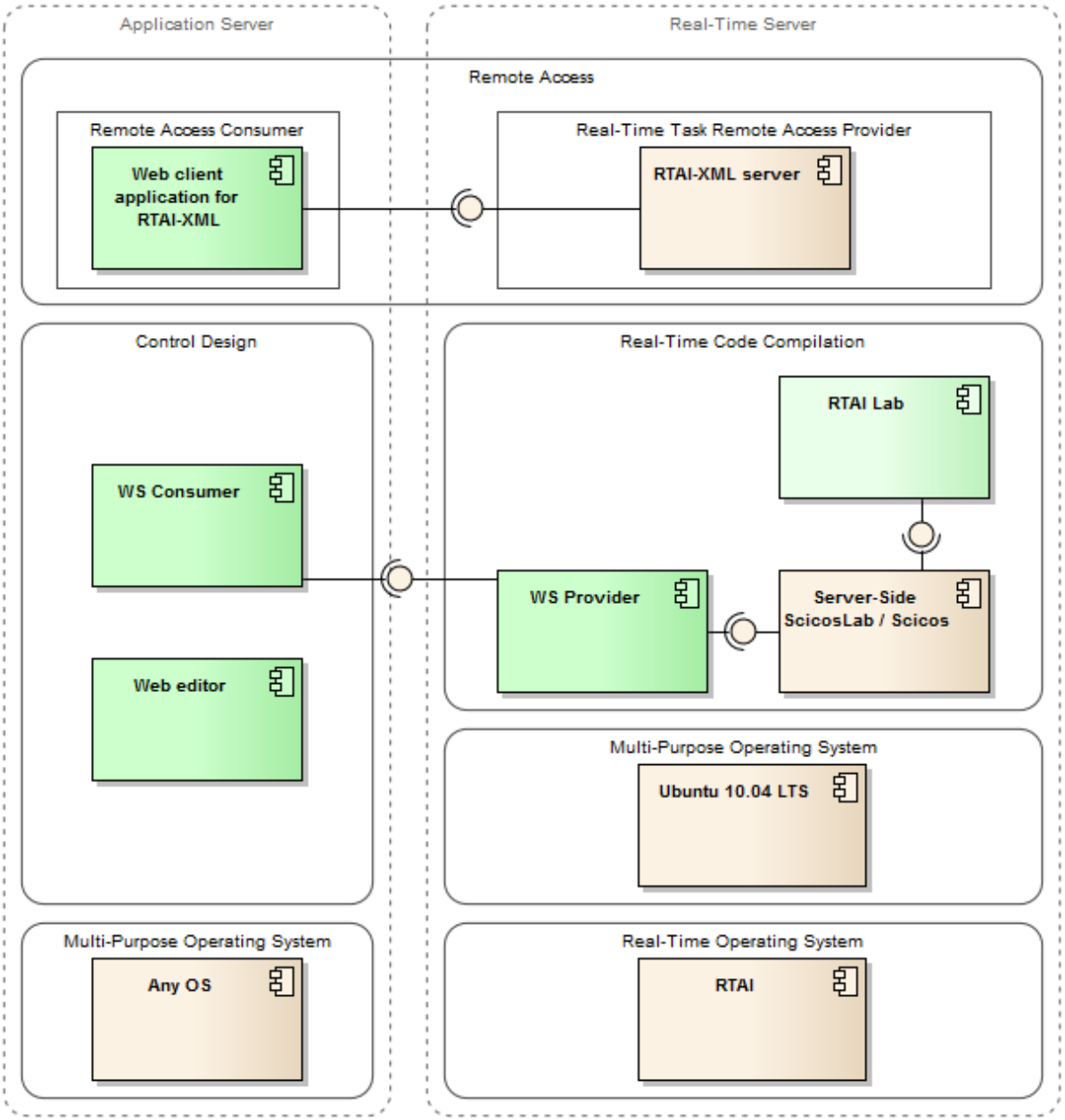


Fig. 1 Architecture diagram of the on-line laboratory environment

ISBN: 978-1-61804-236-1

Framework schemes in our laboratory environment always contain all necessary blocks for generation of signals, measurement and hardware communication, but the core part of the scheme is blank. Thanks to this feature, users are relieved from designing the whole blocks schemes, thus they can focus to the main part of the block scheme (e.g. the controller). The example of such framework scheme can be seen in Fig. 2. Given scheme contains blocks for generating desired value (*w*), closed loop for controller, blocks for limitation of maximum value of action signal to connected hardware, communication with hardware (*Comedi D/A* and *Comedi A/D* blocks), measurement (*scope*), clock signal (input block on the upper right-hand side) and empty superblock marked as *Controller_ID* that is prepared for insertion of custom controller scheme.

The resultant block scheme created by combining the framework scheme and the custom structure of the controller is saved into application's working directory. The WS Provider component executes ScicosLab using a special startup script afterwards. The script loads necessary macros that are required for proper functionality of Scicos, RTAI-Lib toolbox and non-interactive compilation of block scheme into executable binary file. The automated scheme compilation is discussed in [7]. We have configured the RTAI-XML server to allow remote access to the newest real-time task that was compiled using the described method.

Web services provided by WS Provider component are divided into two separate parts: standard user functions and administration. The standard functionality contains following functions:

- *frameWorkSchemeList()* – retuns a list of all framework schemes that are currently available within the WS Provider component on current Real-Time server;
- *frameworkSchemeDescription(schemeName)* – returns a formatted detailed description of the framework scheme whose name was provided in the argument (framework scheme descriptions are managed by administrators using web service described later in this section);
- *compile(frameworkSchemeName, customScheme)* – checks the custom scheme and identifies potentially malicious blocks or code, inserts custom scheme into the framework scheme, executes the scheme compilation and returns ScicosLab's console output;
- *compileTest(frameworkSchemeName,customScheme)* – performs security verification, combines the framework and the custom scheme and returns a status message without executing the compilation step. This function is suitable for developers while integrating the laboratory portal with WS Provider.

For security reasons, we have added a verification procedure and a logging functionality to each provided service. Each user has to be authenticated to be able to execute any of the provided web service successfully. User's credentials have to be provided using HTTP Basic authentication in each SOAP requests. Usage of SSL is strongly recommended regarding the fact that user name and password are sent in open form. Each operation is automatically logged into the database that stores user names, date and time of service call and operation details.

The user and privilege management as well as the framework scheme descriptions can be managed using administrative web service. The administration offers several functions:

- *aclGrant(user, object, operation)* – grants the *user* a privilege to perform *operation* on a selected *object*;
- *aclRevoke(user, object, operation)* – revokes a privilege to perform *operation* on selected *object* from specified *user*;
- *userInsert(user, password)* – inserts new *user* and his *password* into the database;
- *userUpdate(user, password)* – updates *user*'s *password*;

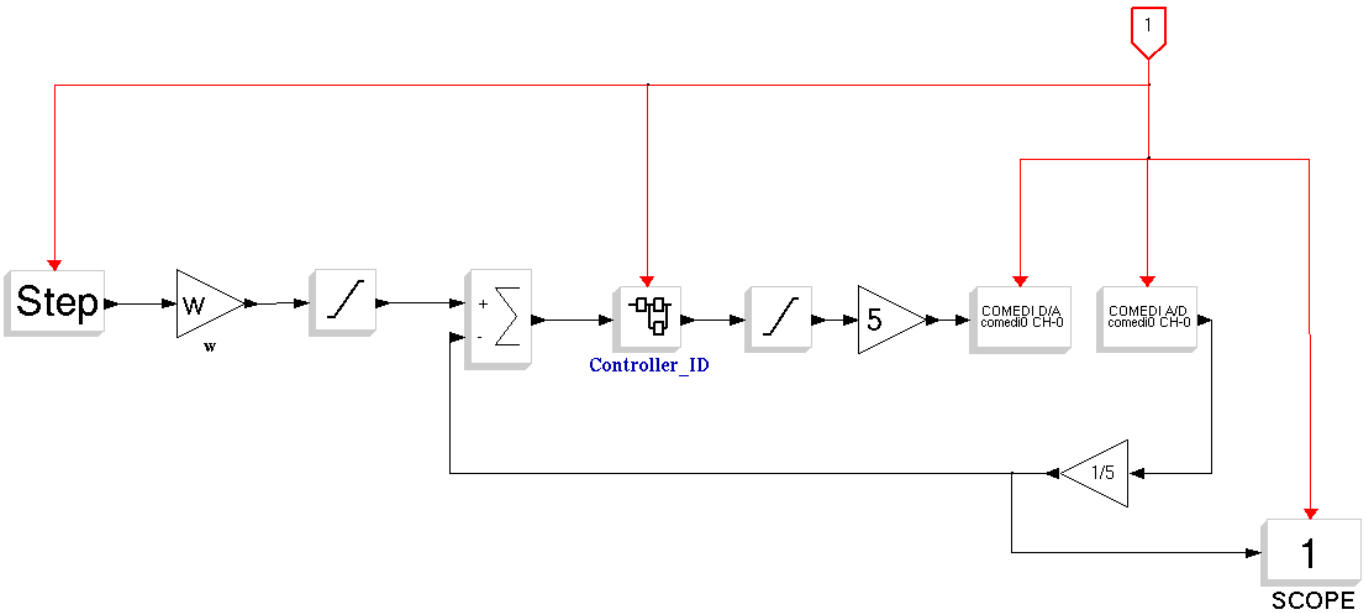


Fig. 2 Framework scheme for magnetic levitation control

ISBN: 978-1-61804-236-1

- *userActivate(user)* – activates *user*'s account;
- *userDeactivate(user)* – deactivates *user*'s account (users may deactivate their accounts also after several failed attempts to call any method using wrong password);
- *frameworkSchemeDescriptionInsert(schemeName, description)* – inserts new version of *description* of the selected *schemeName*;
- *logRead(dateFrom, dateTo, type, user)* – returns application's logs filtered by specified arguments.

Every method mentioned above returns a status message about a success of the operation. User authentication and logging is implemented in the same way as in the standard user web services.

For proper functionality of WS Provider component, it is necessary to use server with real-time environment described in section II. Minimal software requirements are: RTAI-patched Ubuntu Linux, RTAI-XML server and ScicosLab with RTAI-Lib toolbox and compilation macros that are provided together with the WS Provider. The web server and PHP interpreter must be configured to be able to work with HTTP basic authentication and SOAP.

IV. Web Services Consumer

We have also developed a WS Consumer component for communication of client applications with WS Provider. The WS Consumer component can be easily integrated into existing on-line laboratory portals that are intended to be used for custom scheme processing and compilation on the Real-Time server. It may also serve as a sample for developers of custom client applications for the Real-Time server's web services.

The WS Consumer component is not a separate web application as the WS Provider. However, it is a PHP class that is prepared in form of a PHP wrapper. Each function described in previous function is encapsulated inside PHP object methods. These methods are accessible via PHP calls in any web application powered by PHP that includes the WS Consumer class. In this way, the component can be easily integrated into any existing project. The only demand on web server and PHP interpreter is to be able to use PHP's SOAP extension.

V. Conclusion

The hard real-time control experiments were almost exclusively a domain of desktop systems. However, more real-time experiments appear in public on-line laboratories. Unfortunately, they lack certain level of interaction with user since the user can execute usually only a pre-defined experiment with ability to change only a limited number of properties. Thanks to the presented solution, we have managed to enhance possibilities of on-line laboratory system and we have demonstrated a new possible trend in evolution of on-line laboratory systems. Together with the rest of features of the developed on-line laboratory modules portal, this solution could be used as a very flexible support for on-line courses in the area of automation and control.

The described solution is already deployed in our department on the Humusoft CE152 magnetic levitation plant [5]. We are planning to continuously extend the scope of the on-line laboratory, provide more plants and possibly to make public access to the on-line laboratory platform to all interested users.

References

[1] R. Bucher, S. Balemi. (2005). Scilab/Scicos and Linux RTAI – A Unified Approach. IEEE Conference on Control Applications. Toronto, Canada.

[2] A. Gambier. (2004). Real-time Control Systems: A Tutorial. 5th Asian Control Conference. Vol. 2. pp. 1024-1031

[3] A. Guiggiani. (2011). RealTime Suite. Online, http://www.rtaixml.net/realtime-suite/

[4] M. Huba, M. Šimunek. (2007). Modular Approach to Teaching PID Control. IEEE Transactions on Industrial Electronics, ISSN 0278-0046, Vol. 54, No. 6, pp. 3112-3120.

[5] Humusoft, "CE 152 Magnetic Levitation Model website," Online, http://www.humusoft.cz/produkty/models/ce152/, 1991-2014.

[6] Z. Janík, K. Žáková. (2011) Online Design of Matlab/Simulink Block Schemes. In: International Journal of Emerging Technologies in Learning (iJET). ISSN 1863-0383. - Vol. 6, Special Issue 1, pp. 11-13.

[7] Z. Janík, K. Žáková. (2013) A Contribution to Real-Time Experiments in Remote Laboratories. In: International Journal of Online Engineering (iJOE). ISSN 1861-2121. - Vol. 9, Issue 1, pp. 7-11.

[8] Z. Janík, K. Žáková. (2013) One Example of RTAI-Based Remote Experiment. IN-TECH 2013 : Proceedings of International Conference on Innovative Technologies, Budapest, Hungary 10.-13.09.2013. - Rijeka : Faculty of Engineering University of Rijeka, 2013. - ISBN 978-953-6326-88-4. – pp. 273-276.

[9] Z. Janík, K. Žáková. (2014) Performance Tests of MyISAM and InnoDB Database Engines for Online-Based Real-Time Experiments. Cybernetics & Informatics '14. Oščadnica, Slovakia, February 5 – 8. – SSKI, 2014.

[10] Z. Magyar, K. Žáková. (2012). SciLab Based Remote Control of Experiments. 9th IFAC Symposium on Advances in Control Education ACE'12. Nizhny Novgorod, Russia.

[11] M. T. Restivo, J. Mendes, A. M. Lopes, C. M. Silva, F. Chouzal. (2009). A Remote Lab in Engineering Measurement. IEEE Trans. on Industrial Electronics, vol. 56, no.12, pp. 4436-4843.

[12] F. Schauer, M. Ožvoldová, F. Lustig. (2008). Real Remote Physics Experiments across Internet – Inherent Part of Integrated E-Learning. In: Int. Journal of Online Engineering (iJOE), 4, No 2.

[13] I. Zolotová, M. Bakoš, L. Landryová. (2007). Possibilities of communication in information and control systems. Annals of the University of Craiova, Series: Automation, Computers, Electronic and Mechatronic, Vol.4(31), No.2, pp.163-168, ISSN 1841-062.

[14] J. G. Zubía, G. R. Alves. (2011). Using Remote Labs in Education: Two Little Ducks in Remote Experimentation. University of Deusto, Bilbao, ISBN: 978-84-9830-335-3.

Visual attention based extraction of semantic keyframes

Irfan Mehmood, Muhammad Sajjad, Sung Wook Baik*

***Abstract*—**The amount of video data available on the internet and personal devices is increasing exponentially due to revolution of consumer devices, social media and web. To extract the desired information from such a huge video repository in a minimal span of time is a challenging task. Keyframe extraction is an enthusiastic research field that manages video data and provides succinct representation of videos for efficient browsing and retrieval tasks. Various existing keyframe extraction methods utilizes low-level features that results in the loss of semantic details. This paper presents a visual saliency driven framework for keyframe extraction that provides concise versions of video by extracting semantically relevant frames. The proposed visual saliency model helps to bridge the gap between low-level features and high-level information. The visual saliency model is build using static and dynamic saliency maps. The static saliency is derived from color opponent component space using center surround measure. The dynamic saliency is determined using motion intensity and its phase coherence. Then a two dimensional visual saliency curve is estimated by fusing static and dynamic saliency maps. Finally, peak points are calculated in the visual saliency curve that leads to the extraction of the keyframes. Based on different evaluation principles, experimental results demonstrate that the proposed technique successfully extracts semantically significant key frames according to the dynamics of video.

***Keywords*—** Keyframe extraction, visual attention model, video summary evaluation.

I. Introduction

Video is a composite of image sequence, audio tracks, and textual information that conveys information with their own primary elements. Recent growth of video capturing devices, data storage and transmission facilities have resulted into large video libraries. Managing such a huge amount of video data is quite difficult as compared to other forms of media like text and audio. This motivated the researchers to develop systems that are capable of tackling video data in an efficient manner [1]. A basic approach for managing video data is video summarization [2]. Video summarization aims to reduce the amount of redundant data in order to extract the most salient information known as keyframe of the video. The resultant video summary represents the most significant video content. It enables viewers to quickly figure out the important contents of video and help them in searching and retrieval tasks [3]. But to generate a video summary, a full understanding of video is required, which is very difficult for contemporary machines. In video summarization, attaining the effective content of a video is important yet an unexplored aspect. Therefore, it is desired to develop such a framework, which presents the effective elements in video data to the user by considering the semantic information [4].

In literature, many low level approaches have been used for the generation of video summaries but they are not affective as they are inconsistent with the human perception [5, 6]. We want to bridge this gap between low level features and human perception by taking into account the human visual attention. Human visual attention plays an important role in selecting and integrating vital information [7, 8]. An efficient framework designed from the point of view of human perception is provided using a biologically-inspired visual attention model in order to provide semantically relevant summaries.

Irfan Mehmood is with the Digital Contents Research Institute Sejong University Seoul, Korea. (E-mail: irfanmehmood@sju.ac.kr).

Muhammad Sajjad Mehmood is with the Digital Contents Research Institute Sejong University Seoul, Korea. (E-mail: sajjad@sju.ac.kr).

Sung Wook Baik is with the Digital Contents Research Institute Sejong University Seoul, Korea. (E-mail: sbaik@sejong.ac.kr).

* Corresponding author. Tel.: +82-02-3408-3797; fax: 02-3408-4339.

II. Methodology

The bedrock of the proposed framework is the concept of visual attention modeling. Initially the static and dynamic visual saliencies are computed for each frame of the video. Then, an attention value is obtained for each of the visual attention clues. The static and dynamic attention values are fused to obtain an aggregated attention value for each frame. The aggregated attention value of each frame in the video is used to make an attention curve of the video. Finally, the key frames are extracted by finding peak points in the attention curve. Main steps of the proposed framework are shown in Figure 1.

A. Static Visual Attention Model

Consider a video V= {fi; i=1, 2, 3…, N}; where N denotes the total number of frames. These frames are in RGB color space. In human cognitive process, color plays a vital role in the analysis of video contents. Color opponent component (COC) is an efficient color space for improved video perception [9, 10]. To incorporate this cognitive property in our system, RGB color space is converted to COC space.

ISBN: 978-1-61804-236-1

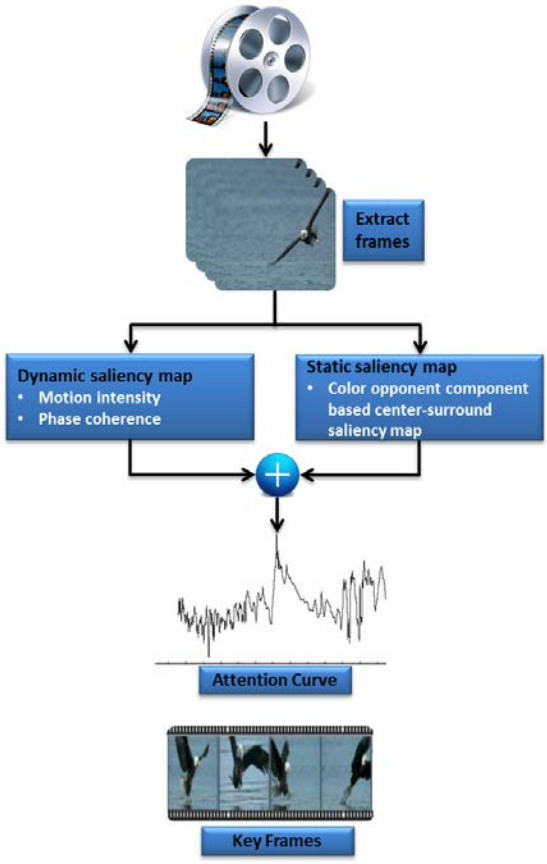


Fig. 1: Framework of the proposed keyframe extraction system

For this purpose, RGB color channels are converted into four-broadly tuned color channels as $R_i=r_i-(g_i+b_i)/2$, $G_i=g_i-(r_i+b_i)/2$, $B_i=b_i-(r_i+g_i)/2$, and $Y_i=(r_i+g_i)/2-|r_i-g_i|/2-b_i$. Then color opponent channels are estimated as:

$$RG_i = R_i - G_i \quad (1)$$

$$BY_i = B_i - Y_i \quad (2)$$

Intensity channel is also computed and fused with red-green and blue-yellow channels to get an aggregated image F_i as given in equations 3 and 4.

$$I_i = \left(\frac{r_i + g_i + b_i}{3} \right) \quad (3)$$

$$F_i = RG_i + BY_i + I_i \quad (4)$$

In this aggregated image F_i, the spatial distribution of color and its spatial position are important factor that contributes to the detection of salient regions in video frames. It has been observed in various studies that salient objects are generally surrounded by non-salient background regions that usually scatter over the entire visual scene [11]. Moreover, objects near the center of the scene are more likely to catch human's attention. These two considerations have led us to the calculation of static saliency map. First, divide F_i into n number of blocks. Now consider an image block B_i, its saliency with respect to other blocks is calculated as:

$$S_S^i = \frac{\sum_{i=1}^{n} e^{1-d_j} \times \frac{1}{D_{i,j}}}{\sum_{i=1}^{n} \frac{1}{D_{i,j}}} \quad (5)$$

here d_j is the Euclidean distance between image block B_i and image's center. $D_{i,j}$ is the distance between image blocks B_i and B_j. This center-surround efficiently estimates saliency map by uniformly highlighting the salient objects. After the computation of the saliency map of the whole image, the average of non-zero values is taken to compute the static attention value of frame.

B. Dynamic Visual Attention Model

In case of videos, human cognitive process is more concerned about objects and motion among them[7, 12]. Therefore, in videos, motion is also a key feature in building human attention model. Furthermore, orientation is an important factor because motion of salient objects is usually associated with consistent orientation. Thus, we have built a saliency model using descriptors motion intensity and phase coherence. Motion vector field V is computed using the Horn-Schunck Method [13]. Motion intensity I_M and phase coherence O are computed from motion vector and are fused to get a dynamic saliency map as:

$$I_M = \sqrt{V_x^2 + V_y^2} \quad (6)$$

$$O = Tan^{-1}\left(\frac{V_y}{V_x} \right) \quad (7)$$

$$S_D = I_M + O \quad (8)$$

where V_x and V_y represents the x and y component of the motion vector V. Final saliency map is obtained by lineally fusing the static and dynamic saliency and normalizing in the range [0 1]. Some results of final saliency maps are shown in figure 2. The average of non-zero pixel values in each saliency map is considered as a saliency value. These saliency values are then computed to from a visual saliency curve. From this saliency curve, keyframes are selected by mapping the peak saliency value to their corresponding video frames.

III. EXPERIMENT AND RESULTS

We have evaluated the proposed summarization framework on videos downloaded from a standard dataset Open Video Project[1], the detail of these videos is shown in table 1. Comparison is done with two different types of video summarization schemes.

[1] http://www.open-video.org/

Fig. 2: First row shows original images and bottom row shows their corresponding saliency maps

Table 1: Details of Test Videos

No.	Video Name	Number of Frames
1	Wetlands Regained, segment 03 of 8	3562
2	Technology at Home: A Digital Personal Scale	3346
3	The Great Web of Water, segment 01	3279
4	The Great Web of Water, segment 02	2118
5	A New Horizon, segment 02	1797
6	A New Horizon, segment 06	1944
7	The Future of Energy Gases, segment 05	3615
8	The Future of Energy Gases, segment 09	1884
9	Drift Ice as a Geologic Agent, segment 05	2187
10	Drift Ice as a Geologic Agent, segment 10	1407

A. Comparison with Non-Visual Attention Based Video Summarization Techniques

Here we have discussed the results of our proposed method with non-visual attention schemes presented in literature. These schemes are STIMO [14] and VSUMM [15]. Method we used for evaluation is based on subjective rating by a group of users. The users are asked to score each video in range [0 100] on the basis of three parameters Informativeness, Enjoyability and ranking. Informativeness measures the ability to maintain all salient content coverage by reducing redundancy; on the other hand Enjoyability measures the performance for selecting perceptually pleasing summaries for video segments. Rank is the overall satisfaction of user with summary contents, scores are assigned for each summary in the range of 0 to 5 and then these ranking score for each summary are averaged to get a single measure, results are shown in table 2. The average results of the three measures indicate that our technique outperforms other mentioned techniques. The average scores assigned by users for Informativeness and Enjoyability for the proposed method are 92.12 and 89.8, respectively that is the highest among competitors. In addition, the score for Rank is also maximum for the proposed scheme. In figure 3, graphical representation of Table 2 have been added to assist the reader in easily understanding the results.

B. Comparison with Visual Attention Based Video Summarization Techniques

In this section comparison is done between the proposed technique with two latest visual attentions modeling based key frame extraction schemes proposed by Peng and Xiaolin [16] and Ejaz et al. [2]. Initially, the results of three techniques are presented on single shot of a video. This video shows an American Bald eagle hunting a salmon on sea surface. The test sequence of frames is from 361 to 605 of the video “Fragment from BBC Nature's Great Events - The Great Salmon Run”. In this shot an eagle dives in the water to catch a salmon and successfully hunt the salmon. In this shot a key frame is one, which shows the eagle flying from surface of sea after hunting a salmon and holding it in his paws. Figure 4 shows the key frames extracted by the proposed scheme, [16] and [2] with frame 512, 384 and 472 respectively. Frame 512 is the key frame conveying more information to user as compared to frames 384 and 472. It is obvious that key frames extracted by other underlying techniques do not express the view of successful hunting of fish by the eagle; thus not semantically representative.

Table 2: Informativeness (I), Enjoyability (E) and Rank (R) score of different methods on video data set

	STIMO			VSUMM			Proposed		
No.	**I**	**E**	**R**	**I**	**E**	**R**	**I**	**E**	**R**
1	74.25	72.75	2.5	80.36	78.5	2.67	92.56	90.5	3.72
2	76.36	70.5	2.1	84.23	80.25	2.7	94.52	90.25	3.8
3	85.25	80.26	2.25	88.32	82.25	2.59	88	92.5	3.75
4	65.32	71.25	2	90	80	2.7	93.33	90	3.8
5	79.36	74.25	2.21	82.36	80.25	2.7	89	87.75	3.7
6	66.25	74.21	2.15	80.25	75.58	2.8	85.62	90	3.6
7	70.3	63.32	2.6	85.26	80.25	2.67	96.6	92.5	3.9
8	65.25	75.25	2.22	80.32	77	2.7	94.75	90	3.8
9	67.25	75	2.24	78.25	77.5	2.9	91.25	85	3.6
10	79.95	74.25	2.48	88.5	77	2.6	95.6	89.5	3.6
Average	72.95	73.10	2.275	83.78	78.85	2.7	92.12	89.8	3.72

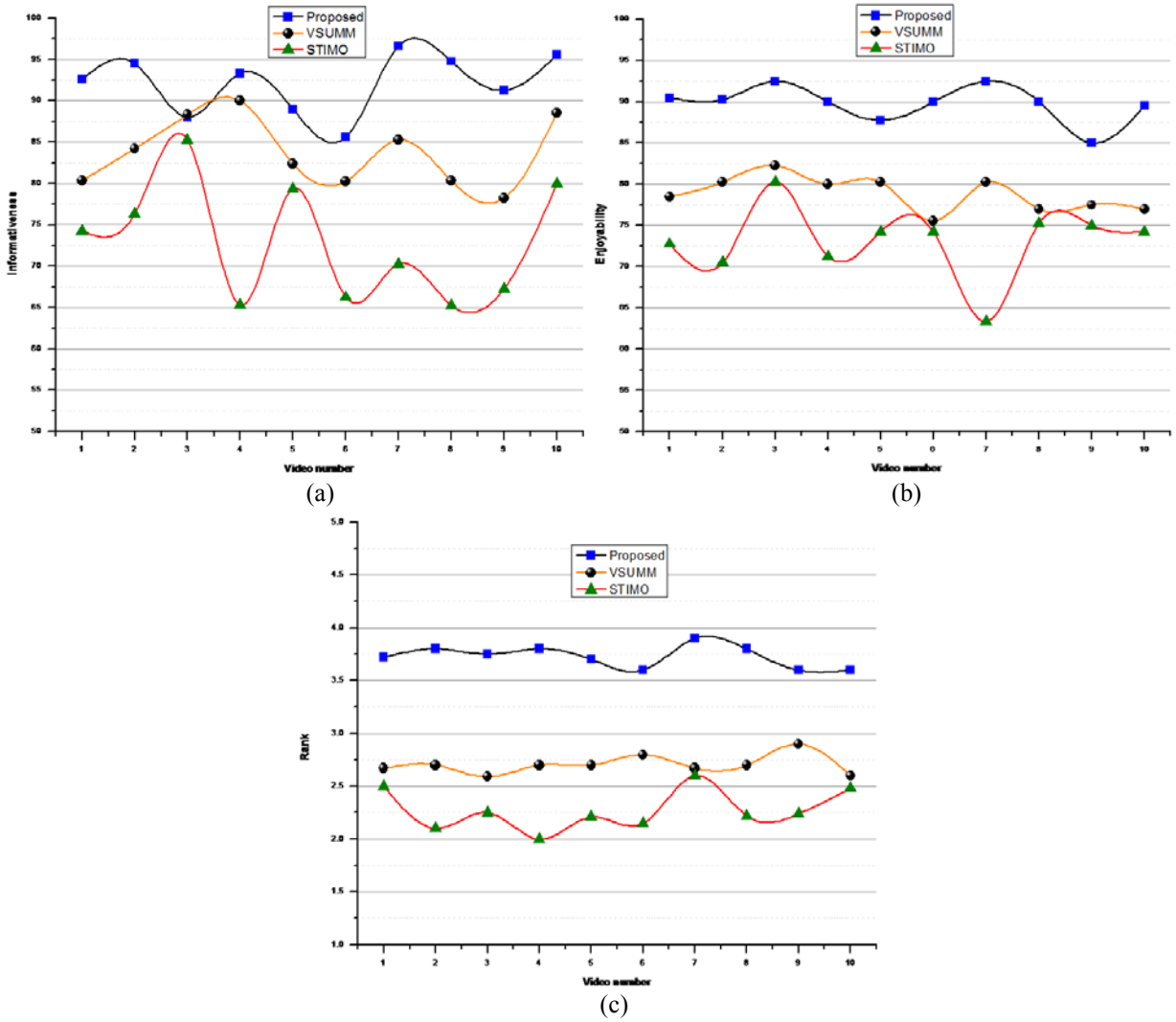


Fig. 3: Informativeness (I), Enjoyability (E) and Ranking (R) curves of different methods on video data set

ISBN: 978-1-61804-236-1

Frame Number 512 Frame Number 384 Frame Number 472

Figure 4: key frames extracted by the proposed scheme, [16] and [2] for the video 'Fragment from BBC Nature's Great Events - The Great Salmon Run'

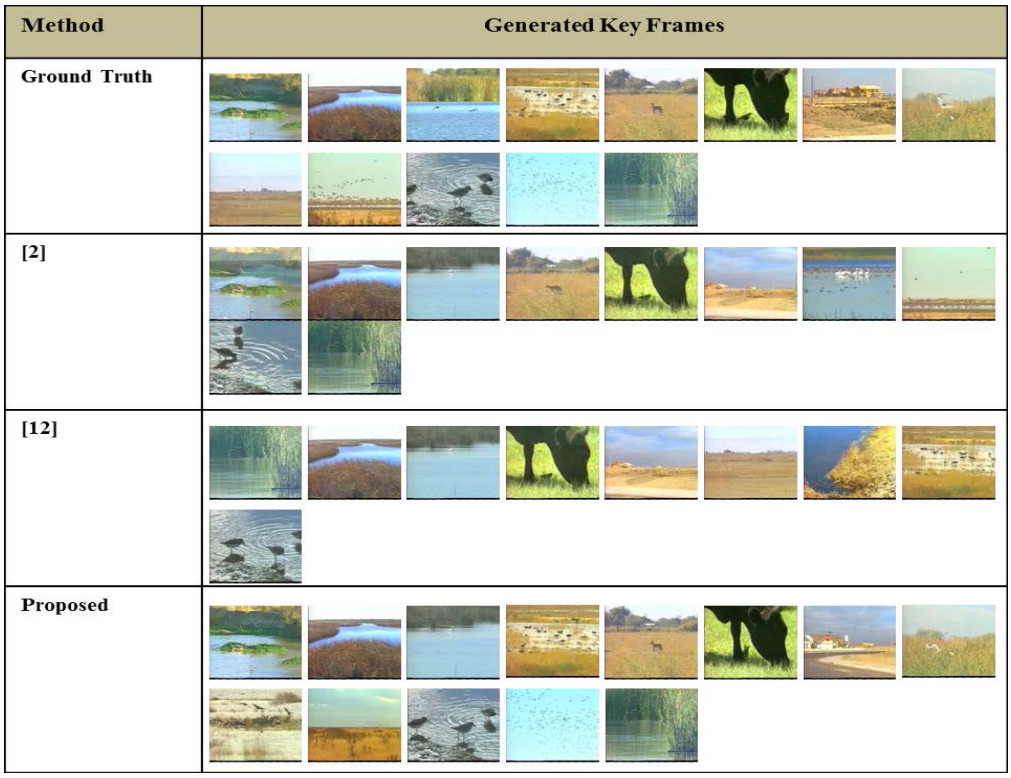


Figure 5: Comparison of key frame extraction for video 'Wetlands Regained, segment 03 of 8'

On the other hand, the key frame extracted by the proposed scheme draw attention and summarizes the shot accurately. Figure 5 shows the key frames for another video 'Wetlands Regained, segment 03 of 8'. This also shows that our scheme yields results closer to the Ground Truth key frames.

IV. CONCLUSION

In this paper, we have presented a method for static and dynamic visual saliency computation and investigate the impending of their fusion for generating videos summaries. This work uses a biological inspired model of saliency which considers different important features such as color contrast, motion intensity and motion orientation between consecutive frames. We believe that the proposed video summarization leads to a more informative and enjoyable summaries for the users. A simple fusion method is used for combining static and dynamic attention values and from this attention curve; key frames are extracted by finding the highest saliency points between two consecutive frames. The proposed evaluation method shows that the extracted key frames are semantically important and closer to the human perceptions as compared to other techniques. In future, we will consider audio features to create video skimming to provide more attractive, natural and informative video summaries.

ISBN: 978-1-61804-236-1

ACKNOWLEDGMENT

This research is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2012904).

REFERENCES

[1] C. Chen, C.-Y. Zhang, Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data, Information Sciences, (2014).
[2] N. Ejaz, I. Mehmood, S. Wook Baik, Efficient visual attention based framework for extracting key frames from videos, Signal Processing: Image Communication, 28 (2013) 34-44.
[3] Q. Xu, Y. Liu, X. Li, Z. Yang, J. Wang, M. Sbert, R. Scopigno, Browsing and exploration of video sequences: A new scheme for key frame extraction and 3D visualization using entropy based Jensen divergence, Information Sciences, 278 (2014) 736-756.
[4] B. Lu, G. Wang, Y. Yuan, D. Han, Semantic concept detection for video based on extreme learning machine, Neurocomputing, 102 (2013) 176-183.
[5] N. Ejaz, T.B. Tariq, S.W. Baik, Adaptive key frame extraction for video summarization using an aggregation mechanism, Journal of Visual Communication and Image Representation, 23 (2012) 1031-1040.
[6] N. Ejaz, S.W. Baik, Video summarization using a network of radial basis functions, Multimedia systems, 18 (2012) 483-497.
[7] M. Guo, Y. Zhao, C. Zhang, Z. Chen, Fast Object Detection Based on Selective Visual Attention, Neurocomputing, (2014).
[8] J.L. Orquin, S. Mueller Loose, Attention and choice: a review on eye movements in decision making, Acta psychologica, 144 (2013) 190-206.
[9] R. Shapley, M.J. Hawken, Color in the cortex: single-and double-opponent cells, Vision research, 51 (2011) 701-717.
[10] L. Dong, W. Lin, Y. Fang, S. Wu, H.S. Seah, Saliency detection in computer rendered images based on object-level contrast, Journal of Visual Communication and Image Representation, 25 (2014) 525-533.
[11] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 33 (2011) 353-367.
[12] S. Spotorno, B.W. Tatler, S. Faure, Semantic consistency versus perceptual salience in visual scenes: Findings from change detection, Acta psychologica, 142 (2013) 168-176.
[13] B.K. Horn, B.G. Schunck, "Determining optical flow": a retrospective, Artificial Intelligence, 59 (1993) 81-87.
[14] M. Furini, F. Geraci, M. Montangero, M. Pellegrini, STIMO: STIll and MOving video storyboard for the web scenario, Multimedia Tools and Applications, 46 (2010) 47-69.
[15] S.E.F. de Avila, A.P.B. Lopes, VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recognition Letters, 32 (2011) 56-68.
[16] J. Peng, Q. Xiao-Lin, Keyframe-based video summary using visual attention clues, IEEE MultiMedia, 17 (2010) 0064-0073.

Irfan Mehmood
received his BS degree in Computer Science from National University of Computer and Emerging Sciences from Pakistan. He is currently pursuing his PhD degree at Sejong University, Seoul, Korea. His research interests include video summarization, prioritization of medical images and brain tumor segmentation.

Muhammad Sajjad
received his MS degree in Computer Software Engineering from National University of Sciences and Technology, Pakistan. He is currently pursuing PhD course in Sejong University, Seoul, Korea. His research interests include super-resolution and reconstruction, Sparse coding, video summarization and mixed reality.

Sung Wook Baik
is a professor in the College of Electronics and Information Engineering at Sejong University. His research interests include Computer vision, Pattern recognition, and Data mining. He has a PhD in Information Technology and Engineering from George Mason University.

A Genetic Algorithm for Shuttering Underperforming Stores

Rong-Chang Chen*, Mei-Hui Wu, Shao-Wen Lien, and Yi-Chen Tsai

***Abstract*—**To earn more profits, the manager responsible for the chain business has to make a hard decision to selectively shutter underperforming stores. How to select to-be-closed stores which do not meet the requirements of the company is a very important issue for the chain businesses. In this paper, a genetic algorithm (GA) is employed to optimally decide which stores are to be closed with the consideration of minimal total revenue loss and minimal total travel distance. To evaluate the efficiency of the proposed algorithm, a number of experiments are performed. Results from this study show that the developed algorithm by this paper can efficiently help select underperforming sites.

***Keywords*—** Store location, Genetic algorithm, Chain business

I. INTRODUCTION

THE location selection of retail stores has long been a focused topic receiving huge attention in retail industry since a few decades ago [1-5]. An appropriate location will bring much money for the company, while an underperforming site might cause some loss in profits [6]. While there is still room for the retail industry to expand, the operating profit is considerably reduced as more and more competitors enter this industry, and thus some operators have to close some underperforming stores to ensure enough profits to be kept.

Two critical factors influence the closing decision. Undoubtedly, the first important factor concerning the closing decision is the revenue. To ensure that the company can make profit, the manager has to close some underperforming stores that no longer meet the company's requirements. The second critical factor to be considered is the travel distance that customers need to drive or walk to get to another store if a specific store is closed. If the required travel distance is too long, customers may choose alternatives from other brands. Therefore, the possible travel distance should be minimized to retain customers. In this paper, we consider these two important factors to develop a two-objective algorithm to help managers to select closing/opening stores.

The remainder of this paper is organized as follows. In Section two, a brief description of the store selection problem is given and the mathematical formulation is presented. In Section three, the proposed approach is briefly introduced. Some experimental results and discussion are presented in Section four. In the final Section, concluding remarks are drawn.

This work was supported in part by National Science Council under grant number NSC 102-2221-E-025-013.

R. C. Chen is with the Department of Distribution Management, National Taichung University of Science and Technology, Taiwan, ROC. (886-4-22196759; fax: 886-4-22196161; e-mail: rcchens@ nutc.edu.tw).

M. H. Wu, S.W. Lien, and Y.C. Tsai are with the Department of Distribution Management, National Taichung University of Science and Technology, Taiwan, ROC. (e-mail: wacs791223@gmail.com; v8200v@gmail.com; angie25567267@gmail.com).

II. THE PROBLEM

To make a clear description, some variables must be defined first. We let $S = \{1, 2, \ldots, s\}$ be a set of chain stores, where s is the number of chain stores that a chain business possesses. Suppose that a percentage p of stores have to keep open. Thus, there will be $s(1-p)$ stores to be closed. Let Q be the number of stores to be closed, where $Q = s(1-p)$ if $s(1-p)$ is an integer; Otherwise, $Q = \text{Int}\{s(1-p)\} - 1$. For $i \in S$, we let $x_i = 0$ if store i is selected to be closed. If not, x_i is set to be 1. Customers have to travel a distance to find another store if a certain store is closed. The longer the travel distance is, the more likely a customer would select alternatives form other brands. To retain customers, the possible travel distance should be as shorter as possible. To evaluate the total travel distance from a closing store i, we define the neighborhood region of store i (as the center) as the region within a radius R, as shown in Fig. 1. Similarly, we define the nearby stores of store i as those stores which are located within its neighborhood region. Since there may exist several stores within the neighborhood region, we designate N_{ij} as the j^{th} nearby store of store i. If the j^{th} nearby store of store i is to be closed, $N_{ij} = 0$; otherwise, $N_{ij} = 1$.The total number of nearby stores for store i is n_i.

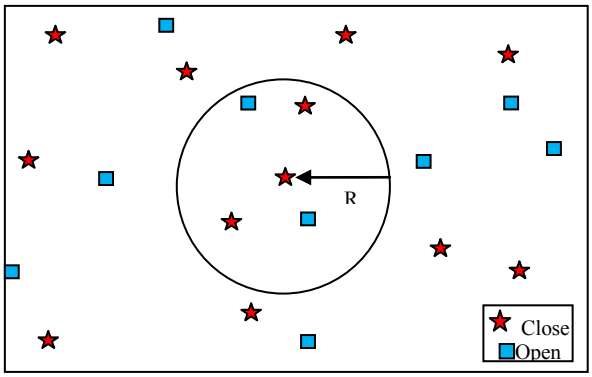


Fig. 1. The neighborhood and the nearby stores of a designated store.

To reduce the customer loss, the total travel distance from closing stores to its nearby stores is minimized. On the other hand, the total revenue loss, which is defined as the revenue difference before and after stores are closed, is minimized. The mathematical formulation of this problem can thus be expressed as:

$$\min Z_1 = \sum_{i=1}^{s} \sum_{j=1}^{n_i} (1 - x_i) d_{ij} N_{ij} \quad (1)$$

ISBN: 978-1-61804-236-1

$$\min Z_2 = \sum_{i=1}^{s} D_i (1 - x_i) \quad (2)$$

subject to

$$\sum_{i=1}^{s} (1 - x_i) = Q \quad (3)$$

$$\sum_{j=1}^{n_i} N_{ij} (1 - x_i) \geq 1 \quad \forall i \in S \quad (4)$$

$$x_i \in \{0, 1\} \quad \forall i \in S \quad (5)$$

$$x_i = \begin{cases} 1 & \text{if store } i \text{ is selected to be open,} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where Z_1 and Z_2 are objective functions, d_{ij} is the travel distance from store i to its j^{th} nearby store, and D_i is the revenue loss of store i if it is selected to be closed.

Equation (3) requires that the number of stores selected to be closed is equal to the required number Q by the company. Equation (4) requires that at least one of nearby stores of store i keep open and this ensures that customers can find at least an open store within the neighborhood. The decision variable x_i is nether 1 or 0, as illustrated in Equation (5). If a store is to keep open, x_i is equal to 1; Otherwise, $x_i = 0$.

When the number of stores grows to be large, the store selection problem becomes very complex. Finding an optimal solution requires huge computation time. Heuristic algorithms, therefore, seem to be more appropriate than exact ones. Amongst the most popular heuristic algorithms, genetic algorithm (GA) [7-20] is one of the most effective and efficient approaches and has been successfully applied in many fields. Some of them are concerned with two-objective problems [18-20]. They employed GA to solve the optimization problems and the results show that GA is quite efficient in solving two-objective optimization problems. We will thus employ GA to solve the present problem.

III. APPROACH

A genetic algorithm based on the concept of a neighborhood of a specific store is developed to solve the selection problem. Since most customers are not willing to walk or drive a long distance to get to another store if a specific store is closed, the utilization of neighborhood concept seems to be appropriate. Within the customer-acceptable neighborhood region of a store which is to be closed, we can keep at least a store open to retain customers.

The GA procedure to solve the problem is illustrated in Fig. 2. The encoding method we used in this paper was a binary scheme. The crossover was done by a uniform method. The evaluation of the fitness functions was performed using a rank-based scheme. The details are described in the followings.

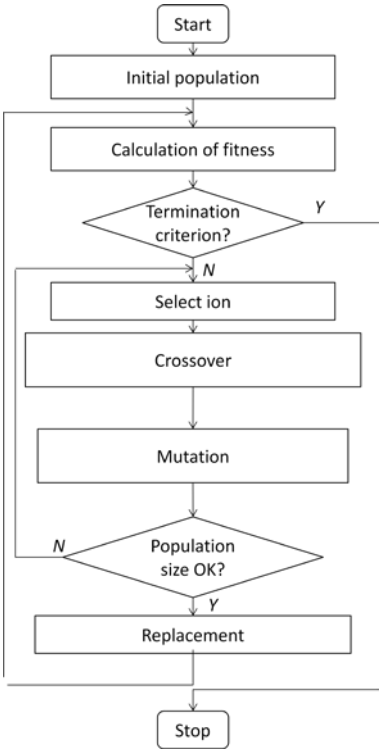


Fig. 2. The GA procedure for solving the problem.

To solve the problem by GA, the chromosome must be encoded first. A binary encoding was employed to represent a chromosome, as shown in Fig. 3. Since the number of stores is s, there are s genes in a chromosome. The value in the gene is either "1" or "0," where "0" means that the store is selected to be closed and "1" means that the store is kept open.

Store	1	2	3	4	5	…	s
Close/open	1	1	1	0	1	…	1

Fig. 3. Representation of a chromosome.

A random method was employed to generate the initial population. The chromosomes were evaluated based on the total revenue deficit and the total travel distance. Nondominated fitness values at each generation were collected and their corresponding solutions were chosen as the Pareto ones. A rank-based approach [18] was utilized to find Pareto solutions. Suppose that a chromosome G_i at generation t is dominated by $p_i^{(t)}$ chromosomes in the present generation. Thus, the rank of the chromosome at generation t can be expressed as [18]

$$\text{rank}(G_i, t) = 1 + p_i^{(t)} \quad (7)$$

All the non-dominated individuals are assigned a value of "1" to their ranks, as illustrated in Figure 4. Take the chromosome marked by a triangle for example. Since there are two chromosomes within the rectangular region indicated by the vertical and horizontal dashed lines, the rank of chromosome is assigned 3. Likewise, the rank of chromosome is 1 if no chromosomes are within its corresponding rectangular region. All the chromosomes

ranked by 1 are collected into the Pareto set. As evolution continues, at each generation the Pareto set is updated and shifts gradually towards exploiting the non-dominated points in the criteria space. The process is continued until the pre-assigned generation number is reached.

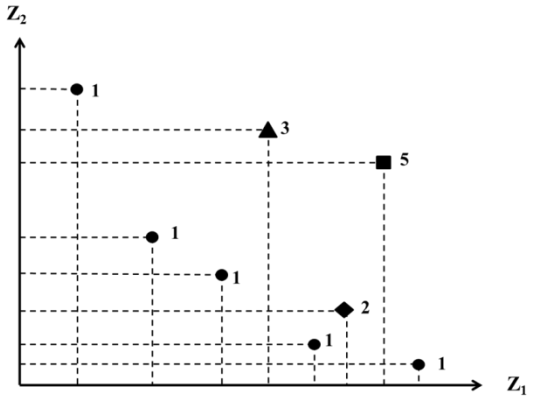


Fig. 4. The ranking in the present scheme. All the non-dominated individuals are assigned rank 1.

The binary tournament selection scheme was used to select fitter individuals. First, two chromosomes are randomly chosen from the population. The fitter one is selected to become a parent A and the process is repeated to select another parent B. Subsequently, GA combines parents A and B to generate the offspring.

In the crossover operation, the uniform scheme was used. In this scheme, a mask is generated randomly. The mask has a same length as the chromosome and is composed of a random set of binary number. Where there is a "1" in the mask, the gene is copied from the first parent; Where there is a "0" in the mask, the gene is copied from the second parent. The offspring, therefore, will contain a mixture of genes from each parent. Figure 5 illustrates the uniform crossover scheme. For example, the first gene value in the mask is "0," meaning that the value of the first gene for the offspring is copied from the second parent. And thus the value will be given "0." The procedure is continued until all the values of genes in the chromosome are copied.

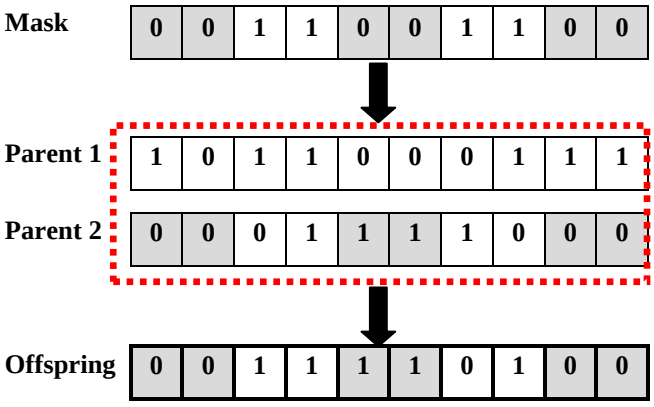


Fig. 5. Illustration of uniform crossover.

In the mutation operation, three genes in the chromosome are randomly chosen and their values are changed, as illustrated in Fig. 6.

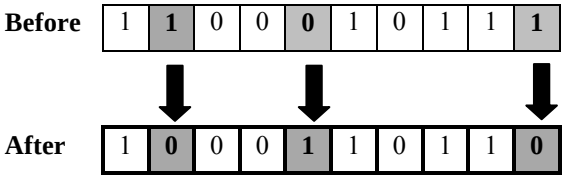


Fig. 6. Illustration of a three-point mutation method.

The GA program will check the number of stores after genetic operations. As the number of stores to keep open is required to be equal to the preassigned value Q, some of the gene values should be changed once the number is different from the required value. If the actual number exceeds Q, some of opening stores will become closing, while some closing stores will become opening if the actual number is smaller than Q. The process is repeated until the number of stores kept open is equal to Q, as shown in Fig. 7.

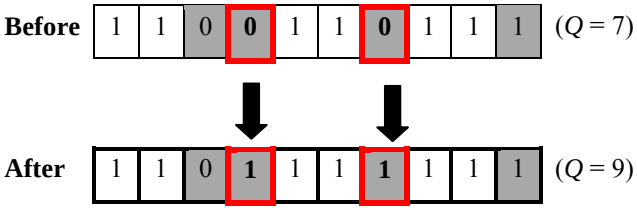


Fig. 7. Illustration of adjustment for the number of stores to keep open $Q = 9$.

The replacement was performed to keep the population size fixed. In addition, poorer chromosomes were replaced with better ones. The procedure is as follows. First, a parent chromosome designated as A is randomly chosen from the population at the previous generation. Then chromosomes at the next generation are compared with the parent chromosome A. If there is any chromosome better than chromosome A, the replacement is done. The process is repeated until all the chromosomes at the next generation are compared.

The program was run until a pre-assigned generation number was reached.

IV. RESULTS AND DISCUSSION

In this paper, we used GA to optimize the store selection. The GA program was coded with Visual Studio C++. The program was run on a Genuine Intel (R) 1.60 GHz CPU and with 1.00 GB RAM. The operating system used was Windows XP. To evaluate the efficiency of the GA program, a variety of experiments were performed using a base case. The dataset was based on a famous store chain in Taiwan. Data were collected from two main cities: Taipei and New Taipei. The base case includes 1429 stores. Their distribution is illustrated in Fig. 8. The neighborhood radius is set to 1,000 m and the generation number is set to be 50,000 at the base case. The percentage of stores to keep open p is set to be 10 %. To faciliate the observation of results, a geographic information system (GIS) [21-22] was employed.

ISBN: 978-1-61804-236-1

Fig. 8. The distribution of chain stores at the base case.

Parts of the optimized results are shown in Fig. 9 and Figs. 10. Figure 9 shows the result for the Taipei and New Taipei cities, while Figs. Shows the results for Taipei city only. The points marked by green color (triangle) are stores to be closed, while those marked by red color (square) are retained to open. Though the store network is quite complex, the GA program developed in this paper is rather easy to use and can help the manager obtain solutions quickly.

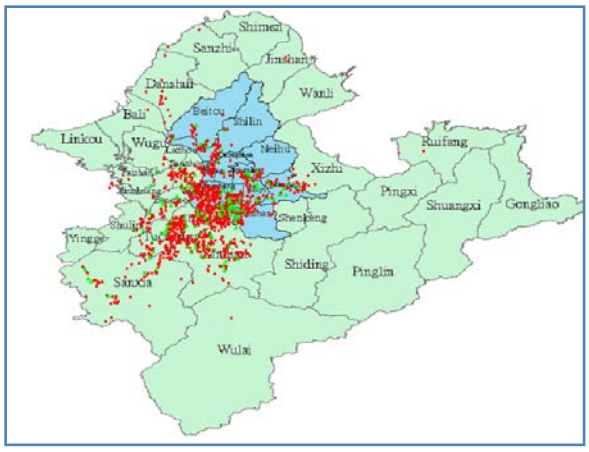


Fig. 9. The result for s = 1429 and R = 1,000 m.

Since most of regions in Taipei city and in New Taipei city are separated by a river or other natural barriers, it is reasonable to divide them into different clusters and make a decision discretely. The results for Taipei city are shown as follows. There are 669 stores in the Taipei city.

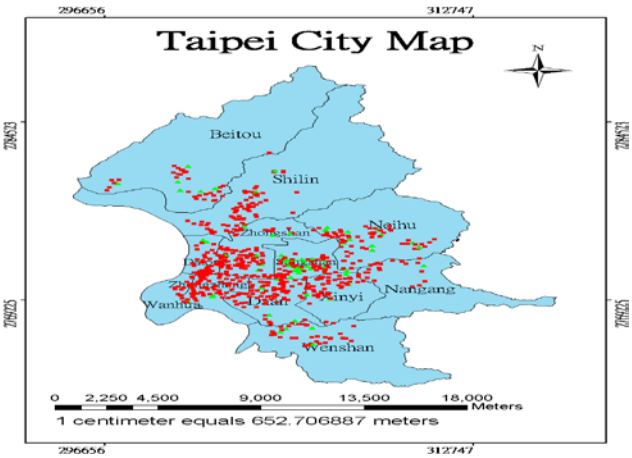


Fig. 10a. The result for s = 669.

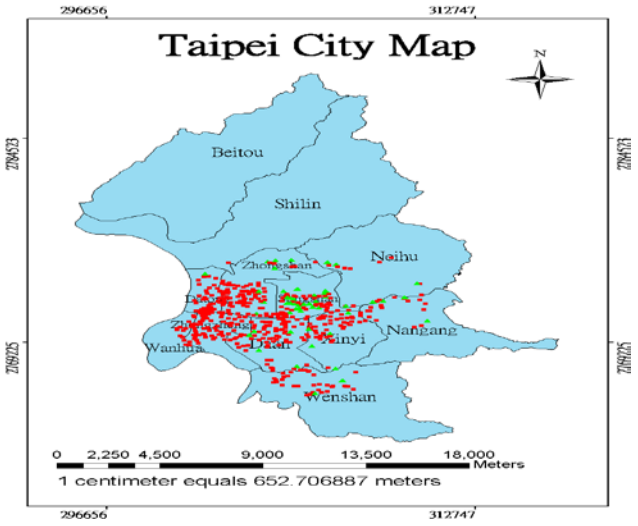


Fig. 10b. The result for s = 500.

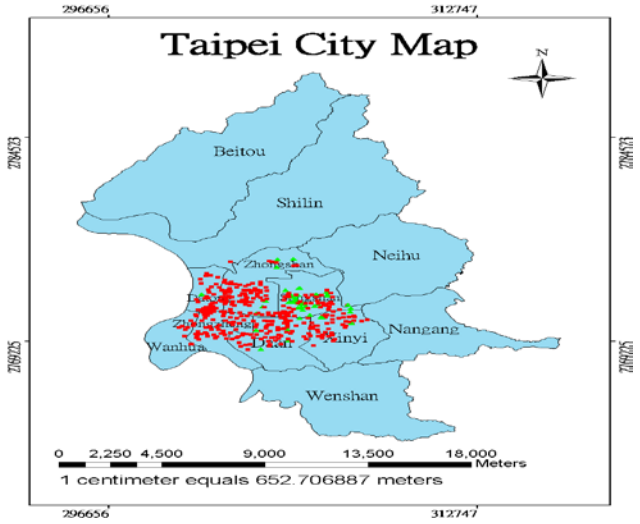


Fig. 10c. The result for s = 400.

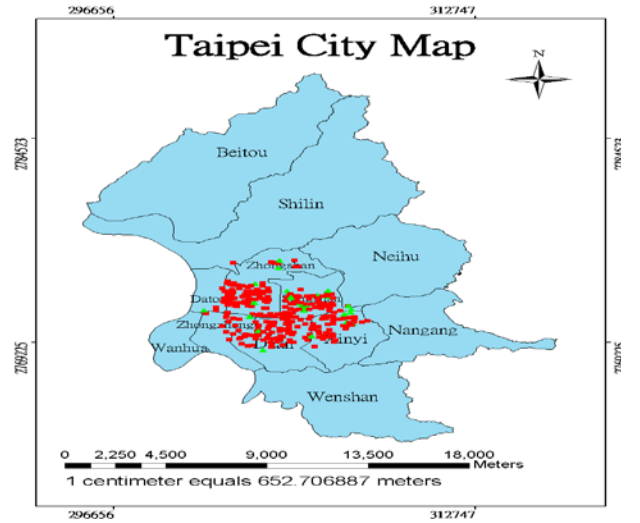


Fig. 10d. The result for s = 300.

As the number of stores increases, the average computation time also increases. Table 1 shows this trend. The results are based on 10 trials. The results show that the GA program we developed in this study is quite

efficient. It takes only about 10 second even if the number of stores is as big as 600.

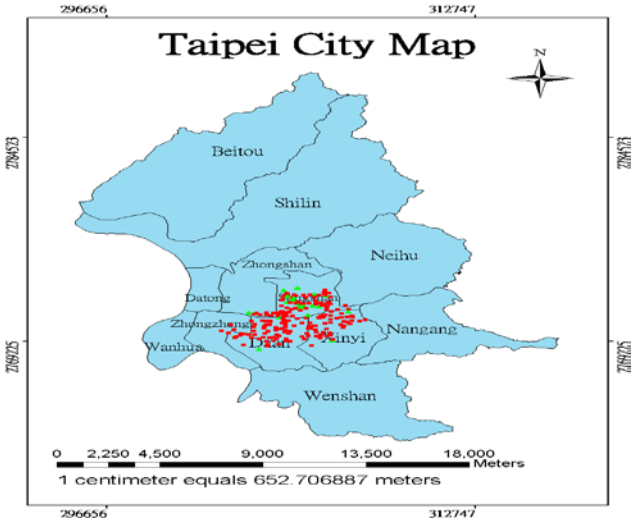


Fig. 10e. The result for s = 200.

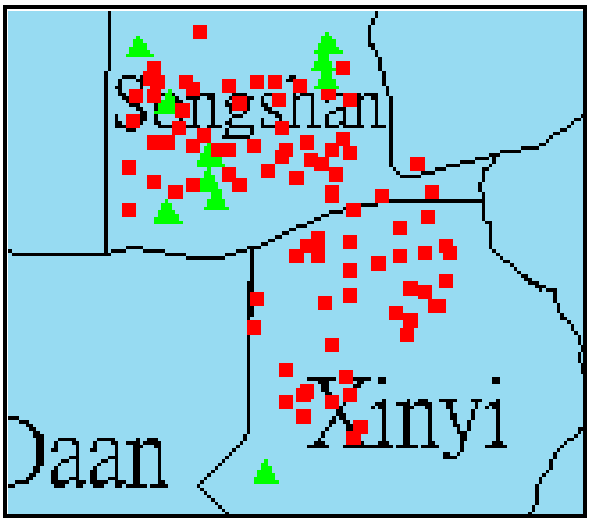


Fig. 10f. The result for s = 100.

Table 1. The variation of average computation time t_{AVG} with the number of stores s.

s	200	400	600	800	1000	1200	1429
t_{AVG}	3.70	7.81	10.55	12.16	22.15	26.77	27.43

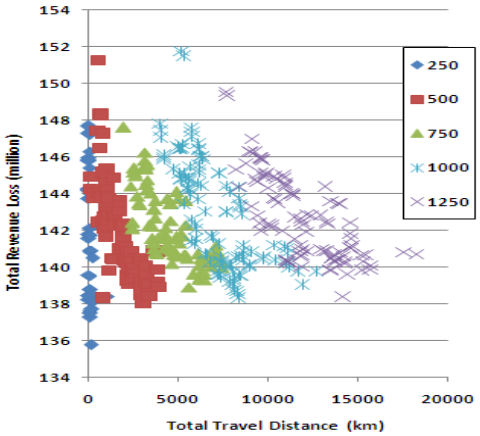


Fig. 11. The variation of total revenue loss with total travel distance at different neighborhood radius R.

The variation of total revenue loss with total travel distance at different neighborhood radius R is shown in Fig. 11. The results were based on 10 trials and their generation number was set to be 50,000. All the Pareto solutions at each trial were recorded. As the total travel distance increases, the total revenue loss decreases. In addition, as the radius increases, the total travel distance increases when the revenue loss is fixed. An interesting finding is that as neighborhood radius increases, the selected stores to be closed gradually move from the city center to suburb where the revenues of stores are lower, as we can compare the results from Figs. 9 and 12.

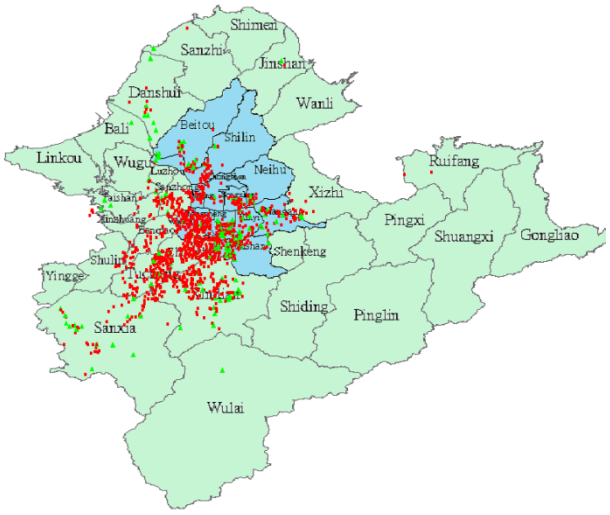


Fig. 12. The result for R = 5, 000m.

V. CONCLUSION

Store chain operators are facing strong pressure as competitors flood the marketplace, and thus their operating profits are significantly reduced. To ensure enough profits, store operators have to think about closing some underperofrmed stores. In this paper, we develop a two-objective genetic alogirthm to help store chain operators to address this issue. To examine the efficiency of the proposed algorithm, some experiments are performed. Results from this study show that the developed GA program can help decide store locations efficiently. In addition, as neighborhood radius increases, the selected stores to be closed gradually move from the city center to suburb where the revenues of stores are lower.

ACKNOWLEDGMENTS

The author wishes to express his appreciation to Shu-Ping Suen, Chih-Chiang Lin, and Mei-Chun Chen for their help during the course of this paper.

REFERENCES

[1] Esichaikul, V. and Suriyalert, C., Site Selection Support Technologies for Convenience Stores. *Vol*., 6 (2010), 19-36.

[2] Vanndelli, K. D. and Carter, C. C., Retail Store Location and Market Analysis: A Review of the Research, *Journal of Real Estate Literature*, 2 (1994), 13-45.

[3] Bai, X., Chen, G., Tian, Q. and Yin, W., Dong, J., Semi-Supervised Regression for Evaluating Convenience Store Location. *International Joint Conference on Artificial*

Intelligence, (2009), 1389-1394.
[4] Tayman. J. and Pol, L., Retail Site Selection and Geographic Information Systems. *Journal of Applied Business Research*, 11 (1995), 46-54.
[5] *R.C. Chen, Y.W. Hsu, Y.H. Ye, and C.W. Huang (2012, Sep). Prediction of Convenience Store Location Based on Support Vector Machines. *International Journal of Digital Content Technology and its Applications*, Vol. 6, No. 16, pp.248-255.
[6] H. Haans and E. Gijsbrechts, "Sales Drops from Closing Shops: Assessing the Impact of Store Outlet Closures on Retail Chain Revenue," Journal of Marketing Research, Vol. 47, Issue 6, pp. 1025-1040, 2010.
[7] J.H. Holland, *Adaptation in Natural and Artificial Systems*, Ann Arbor: University of Michigan Press, 1975.
[8] D.E. Goldberg, *Genetic Algorithm in Search, Optimization, and Machine Learning.* Massachusetts: Addison Wesley, 1989.
[9] M. Gen, and R. Cheng, *Genetic Algorithms and Engineering Design*, Wiley, New York, 1996.
[10] G. Winter, J. Periaux, and M. Galan, *Genetic Algorithms in Engineering and Computer Science*, Wiley, New York, 1996.
[11] M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, 1996.
[12] D.A. Coley, *An Introduction to Genetic Algorithms for Scientists and Engineers*, World Scientific Press, Singapore, 1999.
[13] M. Gen, and R. Cheng, *Genetic Algorithms and Engineering Optimization.* New York: John Wiley & Sons, 2000.
[14] R. Cheng, and L. Lin, *Network Models and Optimization: Multi-objective Genetic Algorithm Approach*. Springer, 2008.
[15] R.C. Chen, M.J. Huang, R.G. Chung, and C.J. Hsu, "Allocation of Short-Term Jobs to Unemployed Citizens amid the Global Economic Downturn Using Genetic Algorithm," *Expert Systems with Applications*, Vol. 38, pp. 7537-7543, 2011.
[16] R.C. Chen, "Grouping Optimization Based on Social Relationships," *Mathematical Problems in Engineering*, Vol. 2012, pp. 1-19. 2012.
[17] R.C. Chen, T.S. Chen, and C.C. Lin, "A New Binary Support Vector System for Increasing Detection Rate of Credit Card Fraud," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 20, No. 2, pp. 227-239, 2006.
[18] C.M. Fonseca and P.J. Fleming, Genetic Algorithms for Multi-objective Optimization: Formulation, Discussion and Generalization, *Proceedings of the 5th International Conference on Genetic Algorithms*, 416-423, 1993.
[19] S.S. Li, R.C. Chen, Y. H. Chen, M.H. Wu, K.H. Leng, and H.Y. Wang, "Application of Multi-Objective Genetic Algorithm to Quotation of Global Garment Companies," *Procedia Computer Science*, Vol. *17*, pp. 173-180, 2013.
[20] R.C. Chen and T.T. Hu, "A Decision-Making Mechanism Considering Carbon Footprint and Cost to Fulfil Orders for Multi-site Global Companies," to appear in *International Journal of Shipping and Transport Logistics*, 2014.
[21] J. Star, and J. Estes, *Geographic Information Systems*. Englewood Cliffs: prentice-Hall, 1990.
[22] M.F. Goodchild, "Geographic Information Systems," *Journal of Retailing*, Vol. 67, No. 1, pp. 3-15, 1991.

Rong-Chang Chen is currently Associate Professor teaching at National Taichung University of Science and Technology (NTCUST). In 1994 he received his doctor degree from National Chiao Tung University (NCTU) at Hsinchu, Taiwan. He has six- year working experience on supply chain management, business administration, and electronic commerce before teaching at NTCUST. His research interests include supply chain management, decision support system, data mining, artificial intelligence, and electronic commerce. Dr. Chen teaches supply chain management and electronic commerce courses. Many of his research works are either published or going to be published in *Neurocomputing, Pattern Recognition Letters*, *International Journal of Shipping and Transport Logistics***,** *Mathematical Problem in Engineering*, *IEEE Transactions on Neural Networks*, *Electronic Commerce Research and Applications*, *Expert Systems with Applications*, *Neural Computing and Applications*, *International Journal of Pattern Recognition and Artificial Intelligence*, *Information-An International Interdisciplinary Journal*, *International Journal of Digital Content Technology and its Applications, Lecture Notes in Computer Science*, *Lecture Notes in Artificial Intelligence*, *WSEAS Transactions on Information Science and Applications*, *WSEAS Transactions on Computers*, *Information Technology Journal*, *IEEE Communications Letters*, *IEE Proceedings Communications*, *Computer Communications*, *ASME Journal of Heat Transfer*, Cryogenics, and more.

.

A method for optimization of plate heat exchanger

Václav Dvořák

***Abstract*—** Research of devices for heat recovery are currently focused on increasing the temperature and heat efficiency of plate heat exchangers. The goal of optimization is not only to increase the heat transfer or even moisture but also reduce the pressure loss and possibly material costs. During the optimization of plate heat exchangers using CFD, we are struggling with the problem of how to create a quality computational mesh inside complex and irregular channels. These channels are formed by combining individual plates or blades that are shaped by molding, vacuum forming, or similar technology. Creating computational mesh from the bottom up manually is time consuming and does not help later optimization. The paper presents a method of creating meshes based on dynamic mesh method provided by software Fluent. Creating of mesh by pulling is similar to the own production process, i.e. it is perpendicular to the plates. The advantages of this method are: The ability to change quickly the whole geometry of the plate, possibility to use optimization algorithms, ability to control the size of the wall adjacent cells and similarity of meshes even in completely different geometries. The paper discusses the problems with very narrow gaps and distortions of the mesh. Using this method, a row of cases with waves and ridges were created. The resulting dependence of efficiency and pressure loss on the ridges count are replaced by mathematical relationships. An objective function is suggested and verified to optimize heat transfer surface of the exchanger.

***Keywords*—**Dynamic mesh, heat exchanger, optimization.

I. Introduction

The development of recuperative heat exchangers in recent years focused on increasing efficiency. Another challenge is the development of so-called enthalpy exchangers for simultaneous heat and moisture transport, i.e. transport of both sensible and latent heat, as presented by Vít et al. in work [1].

A lot of others researchers dealt with design and optimization of plate heat exchangers. For example Gut et al. [2] developed a mathematical model in algorithmic form for the steady simulation of plate heat exchangers with generalized configurations. The configuration is defined by the number of channels, number of passes at each side, fluid locations, feed connection locations and type of channel-flow. The main purposes of this model were to study the configuration influence on the exchanger performance and to further develop a method for configuration optimization.

Fábio et al. in work [3] presented an algorithm for the optimization of heat exchange area of plate heat exchangers. The algorithm was based on the screening method. For each kind of plate, subject to certain constraints, optimal configurations were found which presented the smallest area. Each of these found configurations had local optima characteristics.

Similarly Arsenyeva et al. [4] discussed the developments in design theory of plate heat exchangers, as a tool to increase heat recovery and efficiency of energy usage. The optimal design of a multi-pass plate-and-frame heat exchanger with mixed grouping of plates was considered. The optimizing variables included the number of passes for both streams, the numbers of plates with different corrugation geometries in each pass, and the plate type and size. The mathematical model of a plate heat exchanger was developed to estimate the value of the objective function in a space of optimizing variables. To account for the multi-pass arrangement, the heat exchanger was presented as a number of plate packs with co- and counter-current directions of streams, for which the system of algebraic equations in matrix form was readily obtainable. The exponents and coefficients in formulas to calculate the heat transfer coefficients and friction factors were used as model parameters to account for the thermal and hydraulic performance of channels between plates with different geometrical forms of corrugations. These parameters were reported for a number of industrially manufactured plates. The described approach was implemented in software for plate heat exchangers calculation.

In another work Gut et al. [5] presented a screening method for selecting optimal configurations for plate heat exchangers. The optimization problem was formulated as the minimization of the heat transfer area, subject to constraints on the number of channels, pressure drops, flow velocities and thermal effectiveness, as well as the exchanger thermal and hydraulic models. The configuration was defined by six parameters, which are as follows: number of channels, numbers of passes on each side, fluid locations, feed relative location and type of channel flow. The proposed method relied on a structured search procedure where the constraints were successively applied to eliminate infeasible and sub-optimal solutions. The method can be also used for enumerating the feasible region of the problem; thus any objective function can be used. Examples showed that the screening method was able to

Author gratefully acknowledges financial support by Czech Technological Agency under the project TACR TA01020313.

V. Dvořák is with the Technical university of Liberec, Faculty of mechanical engineering, Department of Power Engineering Equipment. Address: Studentska 2, 46117, Liberec. Phone: +420 485 353 479; e-mail: vaclav.dvorak@tul.cz.

ISBN: 978-1-61804-236-1

successfully determine the set of optimal configurations with a reduced number of exchanger evaluations.

However the optimal design of plates itself is not under investigation in these mentioned studies, while several researchers tried to optimize even the shape of heat exchanger area.

E.g. multi-objective optimization of a cross-flow plate fin heat exchanger (PFHE) by means of an entropy generation minimization technique was described by Babaelahi et al. in study [6]. Entropy generation in the PFHE was separated into thermal and pressure entropy generation as two objective functions to be minimized simultaneously. The Pareto optimal frontier was obtained and a final optimal solution was selected. By implementing a decision-making method, here the LINMAP method, the best trade-off was achieved between thermal efficiency and pumping cost. This approach led to a configuration of the PFHE with lower magnitude of entropy generation, reduced pressure drop and pumping power, and lower operating and total cost in comparison to single-objective optimization approaches.

Kanaris et al. [7] suggested a general method for the optimal design of a plate heat exchanger (PHE) with undulated surfaces that complies with the principles of sustainability. They employed previously validated CFD code to predict the heat transfer rate and pressure drop in this type of equipment. The computational model was a three-dimensional narrow channel with angled triangular undulations in a herringbone pattern, whose blockage ratio, channel aspect ratio, corrugation aspect ratio, angle of attack and Reynolds number are used as design variables. An objective function that linearly combines heat transfer augmentation with friction losses, using a weighting factor that accounts for the cost of energy, was employed for the optimization procedure. New correlations were provided for predicting Nusselt number and friction factor in such PHEs. The authors stated that the results were in very good agreement with published data.

Han et al. in article [8] numerically investigated the thermal-hydrodynamic characteristics of turbulent flow in chevron-type plate heat exchangers with sinusoidal-shaped corrugations. The computational domain contained a corrugation channel, and the simulations adopted the shear-stress transport κ-ω model as the turbulence model. The numerical simulation results in terms of Nusselt number and friction factor were compared with limited experimental data and existing correlations in order to verify the accuracy of the numerical model. The corrugation depth, corrugation angle, corrugation pitch, and fluid inlet velocity were identified as design variables, and 200 samples were selected using the maximum entropy design method to build the metamodel for obtaining the heat transfer coefficient as well as the pressure drop per unit length. A multi-objective genetic algorithm was utilized as the optimizer. The optimization results were presented in the form of Pareto solution set, which clearly showed its dominance over the entire design space and the tradeoff between the two optimization objectives: maximizing heat transfer coefficient and minimizing drop per unit length. Also, the Pareto optimal designs were validated against the values directly obtained from numerical simulations. The approximation-assisted optimization shows that all optimal designs have largest enlargement factor values inside the design space, and the optimal corrugation angle increases with the increase of maximum heat transfer coefficient.

This work is motivated by the need to optimize heat transfer area of plate heat recovery heat exchangers. Heat exchangers are assembled from plates. Plates are made of metal, paper or plastic and are shaped by press molding. Ridges and grooves are supposed to increase the heat (and mass) transfer and also determine the plate pitch and carry the heat exchanger.

To optimize a heat exchanger, we have to create a model and a computational mesh and use computational fluid dynamic (CFD) software. By assembling the heat exchanger, complicated and irregular narrow channels are created. Disadvantages of repeated generation of computational meshes are: It is slow, meshes made in different models are not similar and parameterization of the model is problematic. Further, even a small change of geometry requires to go through the whole process of model creation and mesh generation again. As a result, there is high probability of creation errors of model and low quality of mesh cells. It is necessary to setup the solver, boundary conditions and all models for all computed variants. Furthermore meshes are not similar, i.e. the size, shape, height of wall adjacent cells are not the same for different topologies.

E.g. Novosád in work [9], investigated the influence of oblique waves on the heat transfer surface. The biggest problem in this work was the creation of custom geometry. Each option had to be modeled separately and a meshed. Each model had to be loaded into the solver, set the boundary conditions and subsequently evaluated by calculation.

Therefore, a new method for generation computational variants was developed and is presented in this work.

II. Methods

A. Method for rapid generation of computational model

In this paper, we discuss the case of a counter flow heat exchanger, which has symmetrical heat transfer area. Processes in such heat exchanger can be investigated by modeling the flow around only one plate using symmetrical boundary conditions. How such a model appears can be seen from Fig. 1. The heat transfer surface is divided into two parts. Input and output portions (reported as wall) is fixed, and serve to develop the velocity profiles before the central portion (main wall) which will be deformed. Input boundary conditions are specified by mass flow rates (mass flow inlets), the output boundary conditions are specified as pressure outlets with static pressure 0 Pa.

In this study, we used turbulence model SST κ-ω, medium was air considered as ideal gas. As a results, we obtained pressure, velocity, turbulence and temperature fields inside the computational domain for average inlet velocity of air 3 m/s.

ISBN: 978-1-61804-236-1

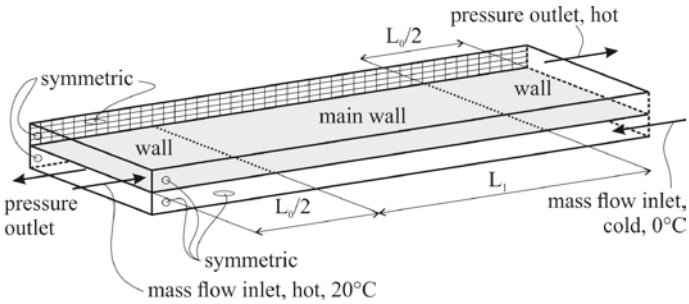


Fig. 1 Model of heat exchange surface of counter flow recuperative heat exchanger.

The new method is based on dynamic mesh method provided by software Fluent. First a simple model with straight wall, see Fig. 1 is created. This model is fully functional, i.e. read into Fluent, boundary conditions are set and it is possible to calculate the flow and heat transfer. In a subsequent step, the computational mesh is deformed using the so-called UDF (User defined functions). Actually, individual nodes of computational mesh are manipulated with. The result is a transformation of the mesh, see Fig. 2 The figure also shows that during the deformation, it is possible to change the spacing of the plates, to create dents or corrugations and also define the size of the cells adjacent to the wall.

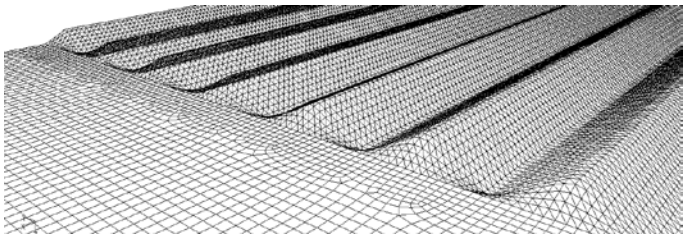


Fig. 2 Computational mesh after deforming.

It is obvious that, as for commonly created mesh, the refinement of the mesh defines minimal size of geometrical details which can be described by the mesh. It can be seen from Fig. 3 that shapes of peaks and valleys of each ridge vary according the grid and ridge spacing.

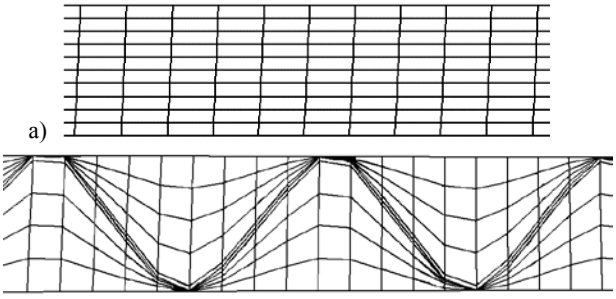


Fig. 3 Cut of the computational mesh, a) – initial mesh, b) – mesh after deforming and ridges creating.

For this initial study, no direct optimization method were used to control the deforming of the surface. Using mentioned method for model generation, we calculated and evaluated two successions of geometries, each with a pitch plates of $\delta = 3$ (mm). In the first case, the geometry of corrugation was defined as waves by function

$$z = \cos\left(\frac{y_0 - y}{2 \cdot y_0}\pi \cdot n\right), \tag{1}$$

where z is relative vertical coordinate of the wall, $2 \cdot y_0$ (m) is width of the model, y (m) lateral coordinate and $n = \langle 0; 30 \rangle$ number of waves. In the second case, the geometry of corrugation was defined as ridges by function

$$z = \arcsin\left[-\cos\left(\frac{y_0 - y}{2 \cdot y_0}\pi \cdot n\right)\Big/(2 \cdot \pi)\right], \tag{2}$$

while tabs of the width $t = 1{,}5\ (mm)$ were created on the tops of the ridges, as it is obvious from Fig. 3.

B. Theory of counter flow heat exchangers

Most of the recuperative heat exchangers in air conditioning works in the isobaric mode, where mass flow rates of warm and cold air are equal, i.e. $\dot{m}_c = \dot{m}_h$. Assuming equality of specific heat capacities, $c_{pc} = c_{ph}$, we can write the coefficient of efficiency as

$$\eta = \frac{t_{hi} - t_{ho}}{t_{hi} - t_{ci}} \cdot 100\ (\%), \tag{3}$$

Where t_{hi} (°C) is the inlet temperature of hot air. Furthermore index c denotes a cold stream, index i inlet into the heat exchanger and index o the outlet of the heat exchanger.

For the pressure drop assessment, it is used local loss coefficient ξ. It is the ratio of total pressure los between the inlet and outlet Δp and dynamic pressure p_d, i.e.

$$\xi = \frac{\Delta p}{p_d} = \frac{\overline{p}_{0i} - \overline{p}_o}{p_d} \tag{4}$$

where Δp is the pressure difference between average total pressure in the pressure inlets $\overline{p}_{0i}$ (Pa) and average static pressure at the pressure outlets $\overline{p}_o$ (Pa).

The dependence between the heat balance and efficiency η is expressed as

$$k \cdot A = \dot{m} \cdot c_p \cdot \frac{\eta}{1 - \eta} \tag{5}$$

where m (kg/s) is the mass flow rate, c_p (J/(kg·K)) is isobaric

ISBN: 978-1-61804-236-1

specific heat capacity, k (W/(m^2·K)) heat transfer coefficient and A (m^2) is the area of heat transfer surface.

III. Results

The results of computations of both cases are plotted in Fig. 4. It is clearly evident from the figure that with the increasing number of waves in the model the efficiency of the heat exchanger increases. The course for both designs of the heat exchanger are similar, but it is evident for smaller count of waves that wall with ridges has higher efficiency than wall with waves. The course of the pressure loss in contrast, is such that for the first few waves, the pressure loss decreases, reaches a minimum for n = 3 ÷ 3.5 and then begins to grow. For comparison, the diagram also shows the results plotted for the wave number n = 0, i.e. for the flat plate. There is clearly apparent advantages of such an arrangement. Variant with waves or ridges needs at least n = 15 to achieve the same efficiency, but the pressure loss is higher comparing to the flat plates.

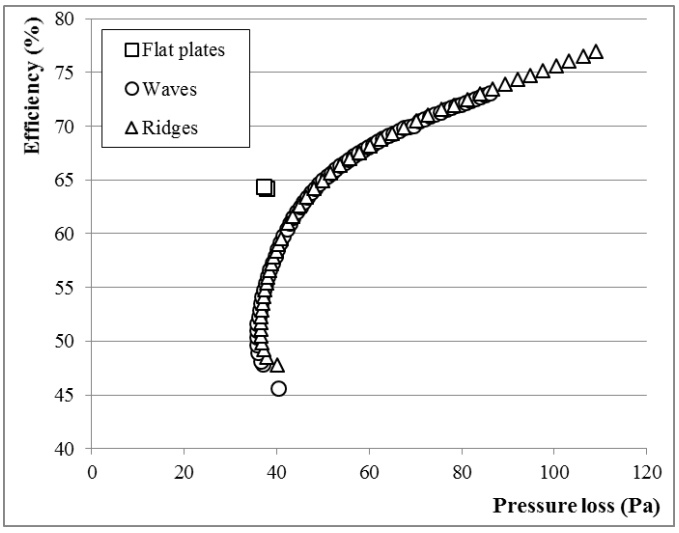


Fig. 4 Diagram efficiency – pressure loss for all variants.

The same results are also plotted in a diagram in Fig. 5, where the heat transfer is represented by heat transfer coefficient k_1 (W/(m^2·K)), which takes into account the heat transfer only in the middle deformed part of the exchanger model and is calculated from the relationship

$$B \cdot (k_1 \cdot L_1 + k_0 \cdot L_0) = \dot{m} \cdot c_p \cdot \frac{\eta}{1-\eta}, \qquad (6)$$

where k_0 (W/(m^2·K)) is the heat transfer coefficient of a flat plate, B (m) is the width of the model, L_0 and L_1 (m) are lengths of flat and deformed parts of the wall respectively, see Fig. 1. Let us mention yet that the heat transfer coefficient is evaluated without respect for the increase in heat transfer area heat caused by deforming. This approach is fully in line with the practice, for which the area of used material is critical.

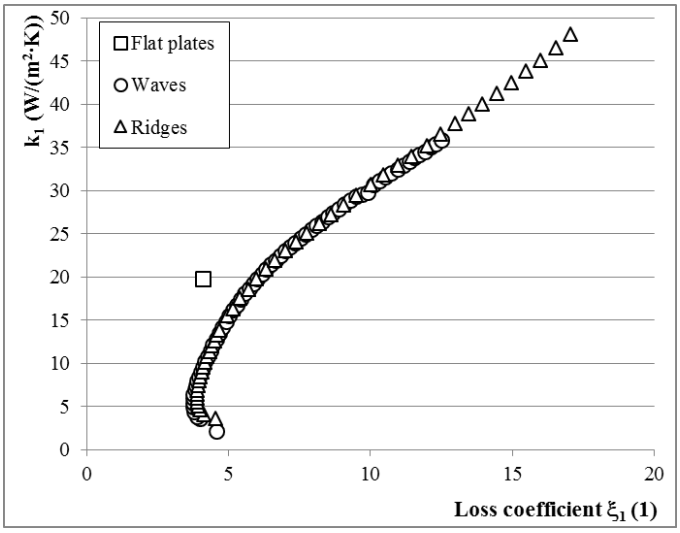


Fig. 5 Surface heat transfer coefficient – pressure loss coefficient for all calculated variants.

Pressure loss coefficient ξ_1 similarly counted the pressure loss only in the deformed part of the heat exchanger surface and is calculated from the relationship

$$\xi_1 = \xi - \xi_0 \frac{L_0}{L_0 + L_1}, \qquad (7)$$

Where ξ (1) is the total pressure loss coefficient of the case and ξ_0 (1) is the total loss coefficient for the case of flat plates.

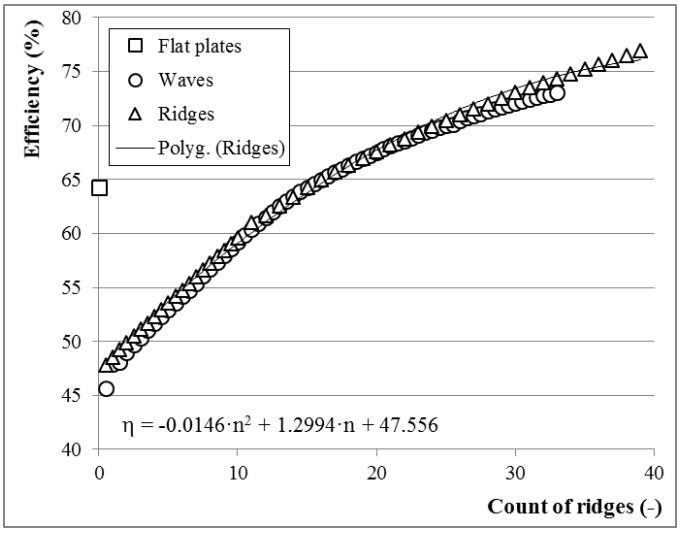


Fig. 6 Dependency of heat exchanger efficiency on count of ridges.

Generally, as we can see from Fig. 4 and Fig. 5, the higher efficiency is obtained for cases with higher pressure loss. However, in practice, pressure loss and efficiency are required when optimizing the heat exchanger. Usually, the maximal pressure loss is specified and for that given pressure restrain, a solution with maximal efficiency is sought. We used obtained dependence of efficiency on count of ridges, which is

presented in Fig. 6, while the dependency of pressure loss is in Fig. 7.

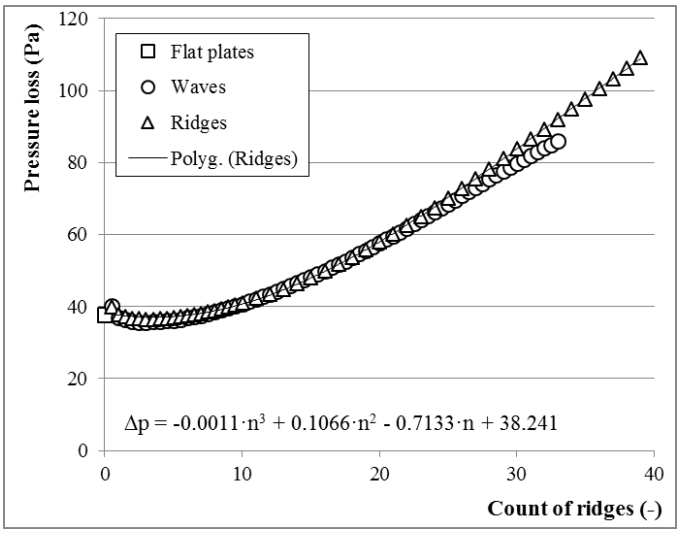


Fig. 7 Dependency of pressure loss of the heat exchanger on count of ridges.

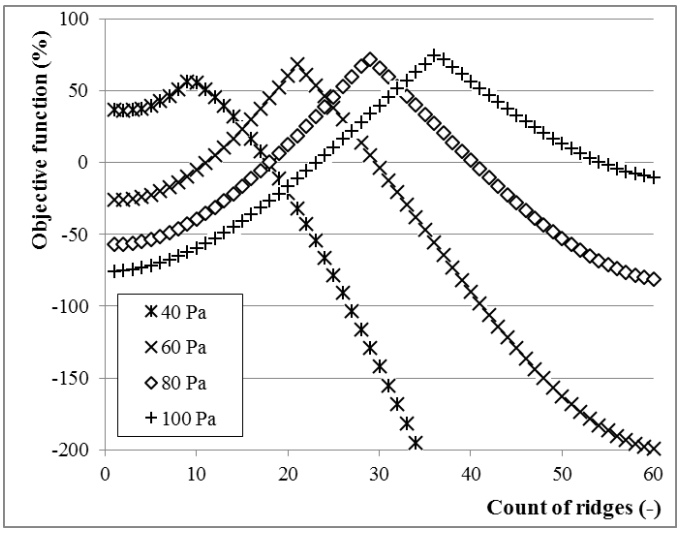


Fig. 8 Course of objective function for various required pressure loss Δp_R.

Because in practice we optimize for maximum efficiency for a given pressure drop, was designed a target function as

$$F = \eta - c_\eta \left(\frac{|\Delta p - \Delta p_R|}{c_E \cdot \Delta p_R} \right)^N , \quad (8)$$

where Δp_R (Pa) is the reference or required pressure loss, c_E (1) is maximal relative deviation of the required pressure drop Δp_R, N is exponent and c_η is penalization of the objective function if the actual pressure Δp loss differs from required. The shape of the objective function was optimized, and it was found that the optimum parameters for the fastest convergence are N = 1, $c_\eta = 2\%$ and $c_E = 0.01$. The course of the target function for required values of pressure loss of the heat exchanger is in Fig. 8.

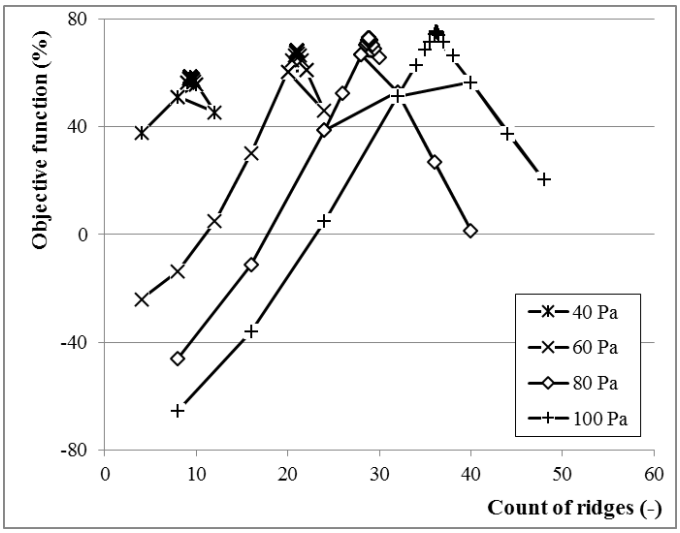


Fig. 9 Course of objective function during optimization for required pressure loss Δp_R.

The courses of objective function during optimization for given required pressure loss, while a simple gradient method was used, are in Fig. 9.

IV. Conclusions

Method has been developed for the rapid generation of computational models. The method simulates the forming process, wherein the heat exchanger plates are made, and allows arbitrarily shaped heat exchanging surface. Among its advantages are the ability to change the pitch of heat exchanger plates and control the size of the cell adjacent to the wall. Another advantage is the similarity of calculating variant and the ability to use the method in the optimization. It is also advantageous to use obtained calculated data for initialization of the next computed variant.

The method was tested on calculation of dependencies of efficiency and pressure drop for two variants of the corrugation of the heat exchange surfaces of heat exchanger. The flow and heat transfer in the heat exchanger with symmetrical heat exchange surface were studied. For forming walls of the heat exchanger, waves and ridges, which proved to be more advantageous, were used. The obtained dependencies were replaced by polynomial functions and an objective function was designed for possible optimization. The function is based on the practice that maximum efficiency of the heat exchanger is required for a given pressure drop constraints. Suitable constants of the function were found and the functions were tested by optimization using simple gradient methods. Only one parameter – number of ridges – was optimized.

The method for rapid generation of computational models and suggested objective function will be used in further work to optimize the shape of the heat transfer surface dependent on more than one parameter.

REFERENCES

[1] T. Vít., P. Novotný, Nguyen Vu, V. Dvořák, "Testing method of materials for enthalpy wheels," *Recent Advances in Energy, Environment*, Economics and Technological Innovation, Paris, France, 29th – 31st October, 2013.

[2] Jorge A. W. Gut, José M. Pinto, "Modeling of plate heat exchangers with generalized configurations," *International Journal of Heat and Mass Transfer 46*, no. 14, 2003, pp. 2571 - 2585.

[3] Fábio A. S. Mota, Mauro A. S. S. Ravagnani, E. P. Carvalho, "Optimal design of plate heat exchangers," *Applied Thermal Engineering*, Volume 63, Issue 1, 5 February 2014, pp. 33–39.

[4] Olga P. Arsenyeva, Leonid L. Tovazhnyansky, Petro O. Kapustenko, Gennadiy L. Khavin, "Optimal design of plate-and-frame heat exchangers for efficient heat recovery in process industries," *Energy* Volume 36, Issue 8, August 2011, pp. 4588–4598

[5] Jorge A. W. Gut, José M. Pinto, "Optimal configuration design for plate heat exchangers," *International Journal of Heat and Mass Transfer* 47, no. 22, 2004, pp. 4833 - 4848.

[6] M. Babaelahi, S. Sadri, H. Sayyaadi, "Multi-Objective Optimization of a Cross-Flow Plate Heat Exchanger Using Entropy Generation Minimization," *Chemical Engineering & Technology*, Volume 37, Issue 1, January, 2014, pp. 87–94

[7] A. G. Kanaris, A. A. Mouza, S. V. Paras, "Optimal design of a plate heat exchanger with undulated surfaces," *International Journal of Thermal Sciences* 48, no. 6, 2009, pp. 1184 - 1195.

[8] W. Han, K. Saleh, V. Aute, G. Ding, Y. Hwang, R. Radermacher, "Numerical simulation and optimization of single-phase turbulent flow in chevron-type plate heat exchanger with sinusoidal corrugations," *HVAC&R Research* 17, 2011, pp. 186 - 197.

[9] J. Novosád, V. Dvořák, "Investigation of effect of oblique ridges on heat transfer in plate heat exchangers, *Experimental Fluid Mechanics 2013*, November 19.-22., 2013, pp. 510 - 514.

A short term user model for Adaptive Search based on previous queries

Albena Turnina

***Abstract*—**In this paper we are extending further the model presented in the paper "A novel approach for modeling user's short-term interests, based on user queries" published in IJCSI [17]. In (Turnina, 2013) we present a novel approach for modeling user short-term interests. The proposed user model is represented by a weighted semantic network, composed of nodes and arcs. Through the model the changes in user's interests would be able to be tracked dynamically. This is done by means of weights assign to nodes and links in the model according to user's searches. In this paper we propose the weighting schema for the model.

***Keywords*—**Adaptive search, Semantic networks, User model, Future Internet

I. Introduction

There are three paradigms for information access of web content in a hypertext environment - searching by browsing, searching via submitted queries and by use of recommendation systems [10]. Personalized search is used to provide individualized collections of pages to a user. More often the search personalization is based on a user model representing the user's interests and preferences [10]. According to (Micarelli et al) the approaches for personalized search falls into several major types, which are: Current Context, Search History, Rich User Models, Collaborative approaches, Result Clustering and Hypertextual Data [10]. Adaptivity is related to the overall ability of the system to modify itself according to a user. It serves to facilitate the activities of a user into a system. Adaptivity can be controlled explicitly by a user or be performed implicitly by the system itself. In that case the activity analysis of a user need to be performed based on his interaction with the system. Adaptivity can be accomplished by a range of methods and approaches, which include collecting statistics for the user, analyzing the knowledge for the user and determining how to adapt the system to the user. The two main strategies for adaptive and personalized search are based on using a search context and personal characteristics of a user. Adaptive search is related to the discovery and use of the context of the search. In that process the current search activity is put in relation with previous activities in order to focus or expand the current search.

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014). Albena Turnina is with the Department of Information Technology of Sofia University Sofia, Bulgaria, Bul. James Boucher 5, phone +359889180 (e-mail: turnina@fmi.uni-sofia.bg)

The personalized information retrieval (PIR) systems can be classified according to different criteria as type of a system and a personalization approach. The type of a system is based on type of service, provided by the system, like Web search systems, Multilanguage search systems, personalized news feeds, e-learning, etc. Personalization approaches fall into three main types: by adaptation of a query, by personalization of search results or by both techniques together. In multilingual systems additional steps for translation to different languages are needed as well [12], [13]. The Personalization of search results is a process of selecting the most relevant results for particular user. This can be performed by number of techniques, such as: pre-ordering of results, filtering of results and result scoring [10]. The techniques for query adaptation include modification of an original query, relaxation of a query or substitution of an original query with one or more adapted queries. The approached for query modification include substitution or expansion of original query with terms or concepts, taken from a reference vocabulary or a user profile, assigning of weights to terms or relations between terms and using the methods of pseudo-relevance feedback and relevance feedback. The modification of a query can be performed explicitly by the user himself as well. In (Liu et al., 2004) approach, the query adaptation is performed by specifying the category of the query. Thus, the system is trying to identify the concepts, related to the query in order to provide the necessary context of the search [9]. The most often the expansion terms are taken from a user profile. These expansion terms serves to represent actual user's preferences and interests and give a context to a search. The exact number of expansion terms used for query modification can be predefined or selected dynamically. (Chirita et al., 2007) argue that the number of the expansion terms can be tuned accordingly to query specifics, such as the length and scope of the query, etc. [4]. The process of a dynamic selection of an expansion terms is known as a "selective query expansion". The main advantage of query adaptation over the other methods for search personalization is the lack of additional processing that can burden the system.

According to (Shen et al., 2005) two main aspects of a personalized search are user's interests and a context of the search (understood as a disambiguation of query). In their system UCAIR they focus on modeling short-term interests of a user, through an approach called eager implicit feedback. In

ISBN: 978-1-61804-236-1

their approach the context of a query is understood as interference of previous queries within the same session over the current query [15].

II. User model

The user profile is essential and the most common part of all personalization systems [1], [7], [11]. The user model represents the user's characteristics and behavior in a system. There are wide variety of approaches for user modeling according to purpose and scope of the system. The user can be modeled according to his demographic characteristics and specific features as level of expertise in a domain, cognitive abilities, visual characteristics and etc. (Brusilovsky, 2001) classification of user's characteristics encompass the knowledge, purpose, background, interests, environment and experience with hypertext of a user [2]. (Gasparini et al., 2011) present a new approach for user modeling in a domain of e-learning, which takes into account contextual user aspects, such as technological, educational, personal, and cultural characteristics [6]. There is variety of techniques and methods for symbolic representation and construction of a user profile. Degree of a complexity of a user profile depends on the purpose of a system. It varies from a simple explicit questionnaire to a complex dynamic structure, containing both explicit and implicit data. Sophisticated user profile has ability to change itself dynamically, reflecting changes of user characteristics and interests. The user profile can take part in a personalization process of Information Retrieval systems in three different ways - when takes place in the retrieval process, when is used for a pre-ordering of search results and when is used for a query adaptation [8]. User profile representation is based on various methods and techniques taken from fields of Information Retrieval and Artificial intelligence. The most popular user profile is keyword profile. Although it is simple and effective it lacks some shortcomings due to its inherent simplicity and inability to represent concepts. Other widely used representation techniques are based on semantic networks or concept taxonomies [8]. Semantic network is a graphical notation able to represent some knowledge by means of nodes and connecting links. According to (Sowa, 1992) the most popular types semantic networks are Definitional networks, Assertional networks, Implicational networks, Executable networks, Learning networks and Hybrid networks [16]. The Learning network has the ability to extend itself dynamically with acquisition of new knowledge. The new knowledge can modify the semantic network in different ways - to add new or remove old nodes and links and to change their weights. The semantic network can be used to represent symbolically the relations between terms and/or concepts and their mutual occurrences in texts or documents. Varies weighting schemas can be implemented to assign weights to nodes and links between them.

The semantic network can model the relation between a word and a concept as well. The mapping of a word to a concept can be accomplished by means of reference vocabulary, by learning mechanism or manually. User profiles, based on semantic network can have different level of sophistication and complexity and can encompass one or more networks. There are number of system in which user profile is represented by semantic networks like WIFS, InfoWeb, SiteIF, ifWeb and PIN. In WIFS system each particular user's interest is modeled by dedicated semantic network. The user model can be represented by ontology as well. A widely deployed ontology user model is an overlay model in which the knowledge for a user is represented as a part of domain ontology. The ontology allows inference over collected facts and drawn of new facts [5]. The user profile can be constructed by set of techniques emerged from various fields such as machine learning, Information retrieval and Artificial intelligence. Nevertheless not very popular, there are approaches based on genetic algorithms and neural networks.

III. The short-term user model based on previous queries

In (Turnina, 2013) we present a novel approach for modeling user's short-term interests based on weighted semantic network composed of two kinds on nodes – term nodes and concept nodes. By proposed model we believe we are able to identify a search context for a particular user in a given moment. We can do so by measuring relatedness between keywords, submitted by a user in his searches [17]. Our approach is based on the idea that there are relations between topics of user's interests, which can be measured, evaluated and used to provide a search context. We propose an approach for user modeling, based on data taken from his previous queries. This data can be collected and used to measure the most significant terms for particular user and their mutual relations. The proposed model would be able to track dynamic changes in user's interests and to represent the most actual topics a particular user is interested in. Our aim is to provide a search personalization, based on a query adaptation. This adaptation can be performed by enrichment of the current query with expansion concepts taken from the proposed model. These expansion concepts are selected from the model based on their weights and weight of a link between them and terms from a current query. The weight of a concept node is cumulative value of weights of all term nodes mapped to it. In this paper we present the weighting schema used to assign weights to concept nodes and to links between concept nodes. We argue that keywords (query terms) used by a user in his searches, reflect his short-term interests and information needs. The user in his searches submits multiple queries, some of which are related to each other. Some of these quires can be a specification of the previous one. Sometimes multiple consecutive queries represent more precisely the user information need than each of them by itself. Therefore, the discovery of such related topics in a series of queries is essential for building short-term user profile. In our work we have been inspired by so-called method of "previous query" used in Google search engine, in which the previous queries submitted by a user impact the search results from the current

query [18]. The proposed model is built by a predefined number of successive user queries as we propose to track 20 queries. We believe that tracking 20 consecutive queries is sufficient to obtain the data needed to populate the model. We also want the model to represent the most current user interests. The model is intended to operate in systems in which the searching is performed by means of keyword. The domain of the model is e-learning environments although it can be implemented in other domain as well. The proposed model is represented by a weighted semantic network composed of two kinds of nodes – one for presenting the concepts (concept nodes) and the other for presenting the query terms (term node). The term nodes are mapped to concept nodes based on a degree of similarity between them via automatic classifier. The concept nodes in the proposed model are predefined and taken from the teacher education ontology. We argue that terms used for searching in e-learning environments will be related to concepts of teacher ontology which models the education domain. The concept nodes in the model are related by weighted links which reflect their usage together in user searches. Concept nodes and term nodes have their own weight, assigned according to proposed methodology. These weights are changed dynamically with user searches as goal of the model is be current and up-to-date with user interests. The weight of a term node reflects the frequency of its occurrence in a series of user queries. The weight of a concept node is a sum of weights of term nodes mapped to it. Thus, the weight of one concept node is proportional to the number of term nodes mapped to it and times of their appearance in user queries. This weight reflects the relative importance of that concept for a particular user. At the same time, the weight of a concept node and the weight of a term node are inversely proportional to the time elapsed since their appearance in user queries. After a user submits a query keywords in its first passes linguistic processing - stemming. Then the system checks whether the terms already exist in the model. If not they are added to a model and mapped to the most suitable concept node. At the same time the default weight is assigned to them and the weight of the respective concept node is increased. In case that particular term node already exists in the model its weight as well as the weight of the respective concept node is increased with a default value. With every new query weights of all nodes in the model are decreased according to a proposed formula. In that way the progressive aging of old topics is performed and eventually they are removed from the model. The proposed weighting scheme is presented below. The gathered information taken from user queries serves to populate the model with his interests. In this way the model is constructed dynamically with only implicit data. Another scenario in which the user initially populates the model explicitly with his preferences and interests can be implemented as well. In this scenario the model will be able to reflect user interests even before user interaction with the system. But that scenario has some shortcoming due to burden the user with need to manually populate the model. Because of this shortcomings we prefer to populate the model strictly with implicit data.

In the process of the initial population of the model the concept nodes are initialized with default weighting value $W0$ equal to 1. After the model is populated with implicit data the weight of the particular concept node will be the sum of this default value and the sum of weights of all term nodes, associated to it.

$$(1)\ Wc = \sum WT + W0$$

The weight of particular node WT is formed on the base of the frequency of its appearance in the user queries. But this value is not static because every time the user submits new query the values of all current terms nodes are recalculated. When the particular term is added for the first time to the model it is initialized with the value of $W0 = 1$.

$$(2)\ Wt0 = a0\ W0 = 1$$

The model is built by a predefined number of successive user queries as we propose the model to track 20 queries. The addition of new term leads to decrease in weight of all of the existing terms in the model with 1/20 (because we track 20 successive queries) according to (2).

In order to reflect the changes of weights of the term nodes we propose to use the formula set out below. The Wti represent the value of the weight of the particular term node after i- subsequent queries.

$$(3)\ Wti = ai\ .Wt0$$

where ai is measured according to formula:

$$(4)\ ai = (1 - 0,05i)$$

where i is equal to a number of elapsed subsequent queries.

The values of i are in the range from 0 for the current query to 20 (because we track 20 successive queries). Thus $a0$ is equal to 1 and this value decreases with each new query and eventually $a20$ becomes 0.

The coefficient ai serves to measure the decay or aging of the particular term. These values are as follows:

$$a0 = (1 - 0) = 1$$
$$a1 = (1 - 0,05) = 0,95$$
……
$$a19 = (1 - 0,95) = 0,05$$
$$a20 = (1 - 1) = 0$$

This on its own turn leads to a recalculation of weights of respective concept nodes.

In case that given term appear k-times in queries the weight of each instance of the term is calculated according to (2).

Then the cumulative value of the particular term is the sum of weights of all its instances.

$$(5) WT = \sum Wt$$

Weights of links between concept nodes measure the degree of relatedness between concept nodes for a particular user. That way the higher the weight of a link between two concept nodes the higher the likelihood that particular user will be interested by a topic represented by two concepts taken together. The weight of a given link is a cumulative value which sums the relations between the given terms in a predefined number of queries (we propose to use 20 successive queries). The proposed weighting scheme is presented below. Relation between two or more terms occurring together in query has value of 1. The value of relation between terms of current query and terms of previous query is set to 0.8. The value of this relation to terms of query previous to previous query is 0.6 and so on. This process can recursively be done to the pre-defined number of previous queries. When two or more terms the link between their concept nodes appears together in the same query is initialized with value of 1. This value is subject of aging according to schema set out below. At the same time the weights are assigned to links between the current term concept node and previous terms concept nodes. But weights of these links will be lower and will decrease with increase of number elapsed queries. In this process, the more distant backward the query is the lower weights of links between its concept nodes and current query concept nodes will be. The intuition behind the model is that the terms occurring together in the same query are the most related and taken together express the information needs of a user. That is why we propose to assign the higher weights to the links between their concept nodes. The information needed for the population of the weights of the links between the concept nodes is gather implicitly and changes dynamically with each successive query. When the model is initialized for the first time the weights of all links are equal to 0. With the information taken from user activity in the system we can build model which reflect the connected topics of user interest. Through the proposed model, we are able to express formally the idea that the relations between query terms exist and they can be measured in the range of queries. These relations will get weaker with the increase of the "distance" between the current query and the previous query. At the same time the model will be able to express the relative importance of some query terms for the user, measuring the frequency of their appearance in his searches. We argue that the proposed model could be able to gather information for creating patterns of a user search behavior. In order to select the expansion concept from the model we propose following approach. First the links between the current query concept and all concepts of model are evaluated. Then three links with higher weight are selected. In second step the weight of respective concepts are evaluated. If they are over the predefined threshold the respective concept is used for expansion. Some heuristics and rules will be needed in order to focus the topic of a query and eliminate the inappropriate expansion concepts.

With each new query, issued by the user, the weights of existing nodes and links will be reduced proportionately. If the value of the weight of some node or link is reduced under predefined threshold, the respective node or link will be removed from the model.

IV. Future work

Future work is aimed at the practical realization and implementation of the model. The heuristics and rules needed for elimination of inappropriate expansion concepts are still subject to investigation. Another research question that needs to be addressed is related to validation of the feasibility of proposed model.

Acknowledgment

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

[1] Brusilovsky, P., Tasso, C.: Preface to special issue on user modeling for Web information retrieval. User Model. User-Adapt. Interact. 14, 147–157 (2004)

[2] Brusilovsky P. (2001). User Modeling and User-Adapted Interaction, 11: 87-110

[3] Callahan, E. (2005). Cultural similarities and differences in the design of university websites. Journal of Computer-Mediated Communication, 11(1)

[4] Chirita, P.-A., Firan, C.S., Nejdl, W.: Personalized query expansion for the Web. In: 30th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007), pp. 7–14. ACM, Amsterdam (2007)

[5] Dolog, P., Nejdl. W,. Semantic Web Technologies for the Adaptive Web. The Adaptive web. Lecture Notes in Computer Science, 2007, Volume 4321/2007, 697-719, DOI: 10.1007/978-3-540-72079-9_23. p.697-719

[6] Gasparini, I., Weitzel, L., Pimenta, M.S. & Oliveira, J.P.M.d. (2011). Adaptive e-learning for all: integrating cultural-awareness as context in user modeling. In T. Bastiaens & M. Ebner (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2011, pp. 1321-1326.

[7] Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A. : User profiles for personalized information access. In: Brusilovsky, P., Kobsa, A., Nejdl,W. (eds.) The AdaptiveWeb, 1 edn, pp. 54–89. Springer, Berlin (2007)

[8] Ghorab, M., Zhou, D., O'Connor , A., Wade, V., Received:Personalised Information Retrieval: survey and classification < http://link.springer.com/article/10.1007%2Fs11257-012-9124-1>

[9] Liu, F., Yu, C., Meng, W.: Personalized Web search for improving retrieval effectiveness. IEEE Trans. Knowl. Data Eng. 16, 28–40 (2004)

[10] Micarelli, A., Gasparetti, F., Sciarrone, F., Gauch, S., : Personalized Search on theWorld Wide Web < http://citeseer.uark.edu/projects/citeseerX/papers/personalized%20search .pdf>

[11] Micarelli, A., Sciarrone, F.: Anatomy and empirical evaluation of an adaptive Web-based information filtering system. User Model. User-Adapt. Interact. 14, 159–200 (2004)

[12] Oard, D.W.: Multilingual information access. In: Encyclopedia of Library and Information Sciences, 3rd edn, Taylor & Francis, Oxford, UK, pp. 3682–3687 (2010)

[13] Oard, D.W., Diekema, A.R.: Cross-language information retrieval. In: Williams M. (ed.) Annual Review of Information Science (ARIST), pp. 472–483. Information Today Inc., Medford (1998)

[14] Reinecke, K.; Schenkel, S.; Bernstein, A. (2010) Modeling a User's Culture. In: The Handbook of Research in Culturally-Aware Information Technology: Perspectives and Models, IGI Global.

[15] Shen, X., Tan, B., Zhai, C.: Implicit user modeling for personalized search. In: 14th ACM InternationalConference on Information and Knowledge Management (CIKM 2005), pp. 824–831. ACM, Bremen (2005)

[16] Sowa, J., (1992) : Buiding a semantic network, < http://www.jfsowa.com/pubs/semnet.htm>

[17] Turnina, A. (2013) "A novel approach for modeling user's short-term interests, based on user queries", International Journal of Computer Science Issues (IJCSI), Volume 10, Issue 1, January 2013, <http://ijcsi.org/contents.php?volume=10&&issue=1>.

A Hybrid Wavelet-Based Distributed Image Compression

S. M. Youssef, A. Abou-Elfarag and N. S. Khalil

***Abstract*—** Due to constraint resources in wireless sensor networks (WSNs), efficient compression and transmission of images have gained wide attention. Distributed image compression and transmission is proposed as a solution in order to overcome the computation energy limitation of individual nodes through sharing processing of tasks and to extend the overall network lifetime, by distributing the computational load among otherwise idle processors. To achieve these goals, the model proposes to split the image into a set of segments of equal size then distribute each segment on separate node to compress and transmit it again which reduces the computational overhead on each node. The proposed model hybrid many algorithms such as Discrete Wavelet Transform, encoding technique called set partitioning in hierarchical trees (SPIHT), binary Code Book (CB) and clustering technique. Good image quality and high peak signal to noise ratio (PSNR) are the two performance indicators should be considered during the image compression and transmission. The simulation results have been validated that the proposed scheme optimizes peak signal to noise ratio and mean-square-error (MSE).

***Keywords*—**Wireless sensor networks (WSNs), set partitioning in hierarchical trees (SPIHT), Code Book (CB) and Distributed Wavelet Transform.

I. Introduction

RECENTLY wireless sensor network (WSN) have drawn the attention of the research community in the last few years since, it reduces costs (installation, time and maintenance costs), increase efficiency and monitor anywhere. A wireless sensor network (WSN) is built of distributed autonomous sensors which composed of few to several hundreds or even thousands of nodes to monitor physical or environmental conditions, such as temperature, sound and pressure and to cooperatively pass their data through the network to a main location. The more modern networks are bi-directional, also enabling control of sensor activity [1].

S. M. Youssef is with the Computer Engineering Department, University Arab Academy for Science and Technology, Alexandria, Egypt (e-mail: sherin.youssef@gmail.com).

A. Abou-El Farag is with the Computer Engineering Department, University Arab Academy for Science and Technology, Alexandria, Egypt (e-mail: abouelfarag@aast.edu).

N. S. Khalil is with the Computer Engineering Department, University Arab Academy for Science and Technology, Alexandria, Egypt (e-mail: nourasamirkh@gmail.com)

Unfortunately, nodes suffer problems such as memory limitations, restricted computational power, energy supplied and narrow bandwidth. These resource constraints form serious difficulties for the design wireless sensor networks [2]. For image-based applications, images have to compress to reduce the size of data (the number of transmitted bits) by removing redundant information such as spatial, temporal and spectral redundancies to save communication energy (power of nodes) because visual data such as still pictures, stream video and monitoring data requires a large amount of information which leads to severe of resource [3]. Distributed image compressing and transmission in a WSN is presented as the solution for the above mention problem and evaluate their performance in terms of energy consumption and image quality in a wireless sensor network.

The advantages of using distributed image compression and transmission in WSNs can be clarified in the following two cases. In the first one, nodes have extremely constrained computation power. Hence, a node does not have sufficient computation power to completely compress a large raw image. In this case, a distributed method to share the processing task is required to overcome the computation power limitation of each single node. In the second one, even if nodes are not extremely computation power constrained, but are battery operated, distributing the computation load of processing every raw image among otherwise idle processors of other nodes extends the overall lifetime of the network [4].

The mechanism proposed in [6] uses a scheme for error robust and energy efficient image transmission over wireless sensor networks. The innovations of proposed scheme are two folds: multiple bit stream images encoding to achieve error robust transmission and small fragment burst transmission to achieve efficient transmission. By uses a scheme based in SPIHT coding of data blocks generated from parent-child relation chips of wavelet coefficients. This parent-child relationship is performed in order to reinforce SPIHT fragilities in bit error transmission cases. The proposed algorithm achieve energy efficient transmission by saving energy consumed on control overhead and device switching from sleep to active. Qin Lu et al. [8] proposed a distributed implementation scheme of the Lapped Biorthogonal Transform (LBT) based on a clustering architecture. They overcome the computation and energy limitation of individual nodes by sharing the processing of tasks. This Approach aimed to prolong the lifetime of the wireless sensor network under a specific image quality requirement [5]. In [7] The

ISBN: 978-1-61804-236-1

proposed algorithm achieve energy efficient compression and transmission of images in a resource-constrained multihop wireless network by distributing wavelet transform processing workload over several groups of nodes along the path from the source to the destination in order to overcome the computation and/or energy limitation of individual nodes and to extend the overall lifetime of the network by distributing the computational load among otherwise idle processors with respect to image quality. Two methods were proposed for exchanging data, the first one is parallel wavelet transform (Divide by rows/columns) and the second one is tilling method.

Simulation results show that scheme optimizes peak signal to noise ratio and mean-square-error (MSE).

This paper is organized as follows. Section II represents the architecture of distributed image compression over a wireless sensor node; section III represents the experimental results; finally section IV represents conclusion.

II. DISTRIBUTED IMAGE COMPRESSION

Before explaining the proposed model there are some terminologies and backgrounds needed to describe image compression as it is related to the proposed model.

A. Background and Terminologies of Image Compression

A common characteristic of most images is that the neighboring pixels are correlated and therefore contain redundant information. Image compression is an application of data compression that aims to encode and reduce the original image with few bits to represent an image by removing the spatial, temporal and spectral redundancies as much as possible to store or transmit data in an efficient form. [19].

As mention before image compression techniques used to compress visual data which requires a large amount of information leads to severe of resource, since WSNs has a limitations in resources, distributed image compressing and transmission used to save resources of WSNs by distributing the workload of task to many groups of nodes along the route from the source to destination which achieve our goal.

B. The Proposed Model Architecture for Distributed Image Compression over a Wireless Sensor Node

Fig. 1 shows the proposed model architecture that has different phases including image splitting phase, sub-region multi-level wavelet, encoding using set partitioning in hierarchical trees (SPIHT), generating binary codebook and transmitting data using clustering technique.

Our contribution in this paper includes using of hierarchical trees (SPIHT) in encoding with the advantage of good image quality, high PSNR especially for color images, it is optimized for progressive image transmission, produces a fully embedded coded file, simple quantization algorithm, fast coding/decoding (nearly symmetric), has wide applications, completely adaptive, can be used for lossless compression, can code to exact bit rate or distortion and efficient combination with error protection [14]. Then integrated with binary codebook generation with the advantage of minimize the average distortion between a given training set.

For more robustness under transmission and to avoid decoding errors, taking the advantage of embedding the Wavelet Transform in distributed image compression and also facilitates progressive transmission of images. Because of the inheritance of multi-resolution nature [9], the benefit from wavelet coding schemes are especially suitable for applications just as scalability, orthogonality, compact support, linear phase and high approximation/vanishing moments of the basis function, efficient multi-resolution representation and embedded coding with progressive transmission [10] , [11].

a) Image Splitting Phase

From source images (512x512), at first, the data is partitioned into n segments R1, R2,., Rn where each segment consists of one or more rows. Second, each node runs one decomposition (1D) wavelet transforms on Ri (i = 1, 2, 3,...n). Once the 1D wavelet transform is completed on all rows, one node collects the intermediate results Q1, Q2,., Qn and divides the results into m segments I1, I2,.,Im. Then each node applies 1D wavelet transform on Ii (i = 1, 2, 3,...m). Finally, a node gathers the 2D wavelet transform results J1, J2,., Jm. This data exchange scheme does not result in any image quality loss compared to the traditional centralized scheme [7].

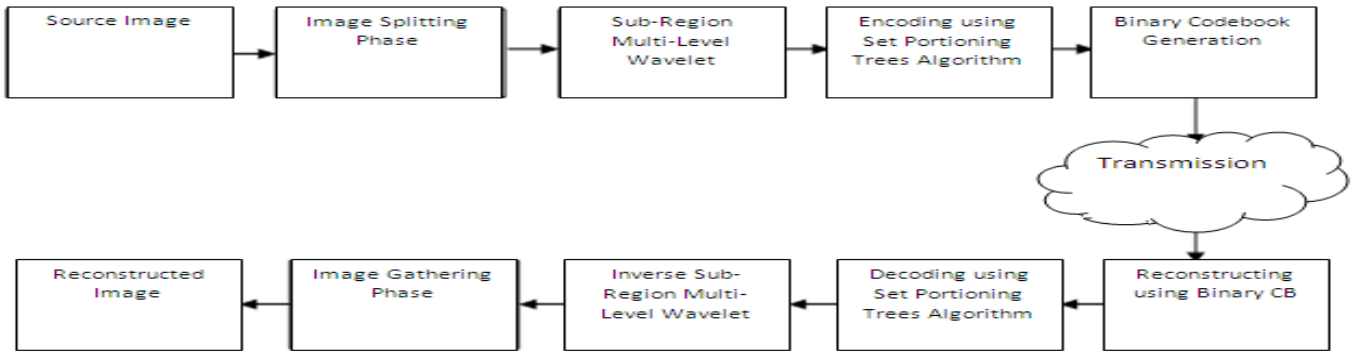
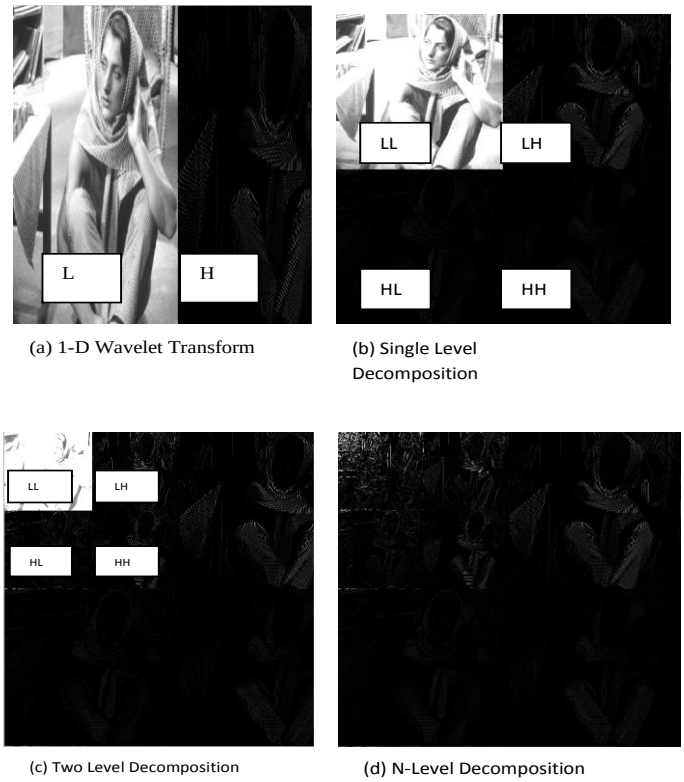


Fig. 1 functional block diagram of architecture of distributed image compression

b) Lifting Scheme for Wavelet Transform (Sub-band Coding)

In recent, wavelet transform has gained widespread acceptance in signal processing in general and in particular in image compression research. The main idea behind wavelet transform is to split up the frequency band of a signal (image in this paper) and then to code each sub-band using octave-band decomposition in our case.

The octave-band decomposition procedure can be described as follows. A Low Pass Filter (LPF) and a High Pass Filter (HPF) are chosen, such that they exactly halve the frequency range of the input signal. First, the LPF is applied for each row of data, thereby getting the low frequency components of the row [2]. The output data contains frequencies only in the first half of the original frequency range because the LPF is a half-band filter. Then, the HPF is applied for the same row of data, and similarly the high pass components are separated. The low and high pass components are arranged into a row of output data as illustrated in Fig. 2(a).



(a) 1-D Wavelet Transform

(b) Single Level Decomposition

(c) Two Level Decomposition

(d) N-Level Decomposition

Fig. 2 illustration of wavelet spectral decomposition and ordering.

This row operation is known as one decomposition (1D) wavelet transforms. Next, the filtering is done for each column of the intermediate output data. This whole procedure including both row and column operations is called a two decomposition (2D) wavelet transform. The resulting 2D array of coefficients contains four bands of data such as LL (low–low), HL (high–low), LH (low–high) and HH (high–high) as shown in Fig. 2(b). The LL band can be further decomposed in the same manner, thereby producing even more sub-bands as shown in Fig. 2(c). This can be repeated up to any level, as shown in Fig. 2(d), thereby resulting in a pyramidal decomposition [2].

c) Set Partitioning in Hierarchical Trees (SPIHT) Encoding Technique

Set Partitioning in Hierarchical Trees (SPIHT) Encoding Technique is primarily a wavelet-based image compression scheme which encodes the decomposed image to a bit stream. In SPIHT algorithm, the wavelet coefficients are fed to the encoder after converted the image into its wavelet transform. SPIHT has been selected because SPIHT and its predecessor achieved better quality when compared to vector quantization and other algorithms. [13].

SPIHT coding operates by exploiting the relationships among the wavelet coefficients across the different scales at the same spatial location in the wavelet sub-bands. Generally, SPIHT coding involves the coding of the position of zero-trees in the wavelet sub-bands and the coding of the position of significant wavelet coefficients [15]. The SPIHT coder exploits the following image characteristics:

1) The majority of an image's energy is concentrated in the low frequency components and a decrease in variance is observed as moving from the highest to the lowest levels of the sub-band pyramid
2) It has been observed that there is a spatial self-similarity among the sub-bands, and the coefficients are likely to be better magnitude-ordered if moving downward in the pyramid along the same spatial orientation [15].

To describe the spatial relationship on the hierarchical pyramid, a tree structure, termed spatial orientation tree shown in Fig. 3 shows how the spatial orientation tree is defined in a pyramid constructed with recursive four-sub-band splitting. Every pixel in the image signifies a node in the tree and is determined by its corresponding pixel coordinate. Its direct descendants (offspring) symbolize the pixels of the same spatial orientation in the next level of the pyramid. The tree is defined in such a manner that each node has either no offspring (the leaves) or four offspring's, which at all times form a group of 2 x 2 adjacent pixels. In Fig. 3, the arrows are directed from the parent node to its four offspring's. The pixels in the highest level of the pyramid are the tree roots and are also grouped in 2 x 2 adjacent pixels. Nevertheless, their offspring branching rule is different, and in each group, one of them (indicated by the star in Fig. 3) has no descendants [16].

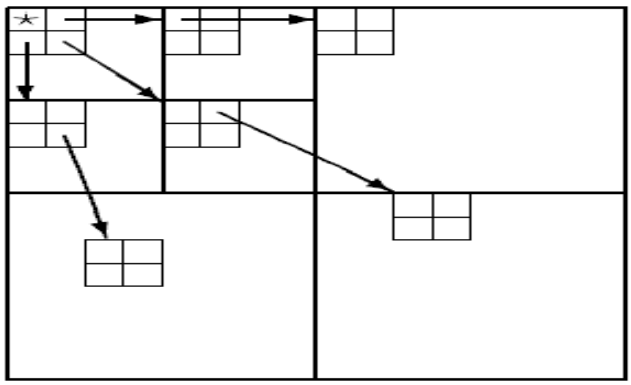


Fig.3 spatial-orientation trees

d) Codebook (CB) Generation

Before explaining binary codebook generation there are some terminologies and backgrounds needed to describe recent algorithms that used to generate codebook such as vector quantization.

1) Vector Quantization (VQ)

Vector Quantization (VQ) is efficient and simple approach for data compression, because it is simple and easy to implement, VQ used to generate a good codebook such that the distortion between the original image and the reconstructed image is the minimum. In the past years, many improved algorithms of VQ codebook generation approaches have been developed.

VQ is a mapping function which maps k-dimensional vector space to a finite set CB = {C1, C2, C3… CN}. The set CB is called as codebook consisting of N number of codevectors and each codevector Ci = {ci1, ci2, ci3… cik} is of dimension k. Good codebook should be designed to reduced distortion in reconstructed image. For encoding, image is split in blocks and each block is then converted to the training vector Xi = (xi1, xi2… xik). The codebook is looked for the nearest codevector Cmin by computing squared Euclidean distance [12].

2) Binary CodeBook (CB) Generation

Due to the output of SPHIT is a stream of bits, binary codebook (CB) could be generated by dividing the output matrix from SPHIT into n*n blocks to generate training vectors matrix then generate binary CB (truth-table) with size of 2^n then match between training vectors and CB to get indexed matrix, suppose n = 2 so CB size will be 4 instead of sending CB with size 128 or 512 or 1024 CB with size 4 will be sent as shown in Table I.

Table. I codebook with size n = 2

CodeBook		
Indexed	Codeword's	
1	0	0
2	0	1
3	1	0
4	1	1

III. EXPERIMENTAL RESULTS

The proposed model has been carried out on several experiments to test the efficiency of the model. The model has been tested using different benchmark images. Moreover, different results from the scheme have been compared with results of other algorithms in [12, 15, and 17].

In this section the performance indicators used to evaluate the proposed model is illustrated.

A. Performance Measures

In order to properly evaluate the performance of image compression schemes and to allow a fair comparison between different schemes, a benchmark suite must include a set of tests and a way of measuring the results of the tests using controlled conditions.

In compression, the tests are oriented to measure the requirements of an application. So, the robustness, the fidelity and the capacity are commonly measured.

a) Mean-Square-Error (MSE):

MSE is the cumulative squared error between the original and reconstructed image as shown in (1)

$$MSE=\frac{1}{M*N}\sum_{x=1}^{M}\sum_{y=1}^{N}[\,I(x,y)-I'(x,y)]\quad(1)$$

Where M, N is the dimension of Image, I(x, y) is the original image, I' is the reconstructed image.

b) Peak-Signal-To-Noise-Ratio (PSNR):

PSNR as shown in (2) is one of the most common measures of distortion in the image field. PSNR is a useful tool to measure perceptibly level, it not always accurate to human eyes adjustment.

$$PSNR=10\log_{10}(255^2/MSE)\qquad(2)$$

c) Compression Ratio (CR):

CR. as shown in (3) is the ratio of the original (uncompressed) image to the compressed image.

$$CR=\frac{Uncompressed\ Image\ Size}{Compressed\ Image\ Size}=\frac{Usize}{Csize}\qquad(3)$$

Where Usize = M x N x K and Csize = size of compressed image file stored in a disk. Where M, N is the dimension of Image, K is 8 bit image.

B. Results

Different experiments have been carried out to test and validate the proposed model using the different data sets. The performance indicators described above were used to provide statistical evaluation of performance without effect of noise. Examples of the tested images are shown in Fig. 4.

Fig.4. examples of the tested images

The PSNR and MSE were calculated for different images using different algorithms as shown in Table. II which clarify that PSNR and MSE of proposed algorithm is better than PSNR and MSE of other algorithms in [12, 15, and 17].

Table. II values of the PSNR and MSE for different images under different algorithms

Tested Images	CB-Size	LBG		REV		Proposed Algo.	
		PSNR	MSE	PSNR	MSE	MSE	PSNR
Tiger	128	21.2	491.4	22.8	340.0	11.9	37.4
	256	21.3	487.7	23.5	288.7		
	512	21.3	480.5	24.4	235.5		
	1024	21.5	465.7	25.2	195.2		
Straw-berry	128	18.4	933.5	23.9	266.8	21.6	34.8
	256	18.5	925.9	24.6	228.2		
	512	18.5	912.8	25.4	186.7		
	1024	18.7	885	26.3	152.9		
Taj-mahal	128	18.5	910.2	23.3	301.2	11.5	37.6
	256	18.6	902.7	24.3	241.3		
	512	18.6	889.4	25.6	179.1		
	1024	18.8	862	26.7	137.6		
Gan-esh	128	20	650.9	21.3	481.6	42.4	31.9
	256	20	645.6	21.9	421.7		
	512	20.1	635.2	22.6	354.7		
	1024	20.3	613.3	23.3	307.9		
Scen-ary	128	22.6	356	25.3	191.3	1.46	46.4
	256	22.7	353	26.3	153.1		
	512	22.7	346.5	27.4	119.1		
	1024	22.9	333.8	28.6	90.2		

Compared with LBG and REV the proposed model in Fig. 5 shown an outstanding improves in performance.

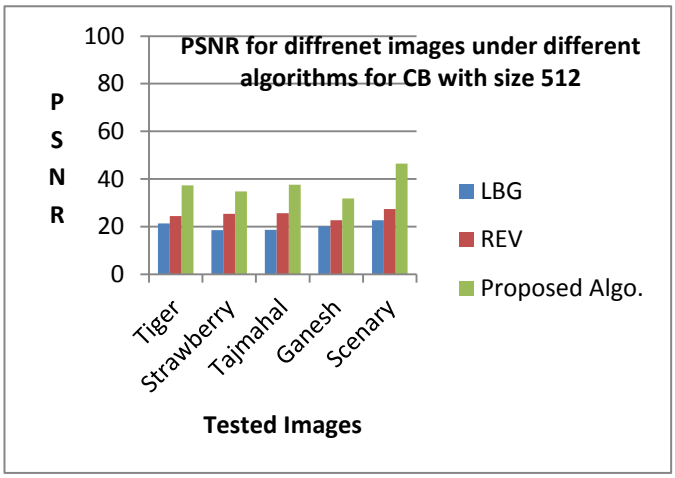


Fig. 5 PSNR for different images under different algorithms for CB with size 512

The dataset images have applied under different algorithms (Two-Level KPE, KPE and proposed algorithm). Table. III demonstrates changing of the PSNR and MSE for different images under different algorithms.

Table. III values of the PSNR and MSE for different images for different algorithms

Test Images 512x512		Two-Level KPE		KPE		Proposed Algorithm	
	CB Size	MSE	PSNR	PSNR	MSE	MSE	PSNR
Tiger	256	267	23.9	21.5	463.8	11.9	37.4
	512	131	26.9	21.8	432.2		
	1024	68.5	29.8	22.7	345.7		
Straw-berry	256	168	25.9	22.8	338.1	21.6	34.8
	512	87.1	28.7	23.7	277.3		
	1024	49.9	31.2	24.5	233.6		
Taj-mahal	256	159	26.1	22.5	364.6	11.4	37.6
	512	77.7	29.2	23.7	279		
	1024	46.3	31.5	24.5	230.3		
Gan-esh	256	307	23.3	20.0	610.6	42.4	31.9
	512	204	25	20.0	533.1		
	1024	125	27.2	20.8	449.1		
Scen-ary	256	117	27.5	22.0	406.4	1.5	46.5
	512	73.1	29.5	23.4	296.8		
	1024	34.4	32.8	25.3	189.2		

As shown in Fig. 6, the PSNR of our proposed model remains high compared to Two-Level KPE and KPE for all tested image.

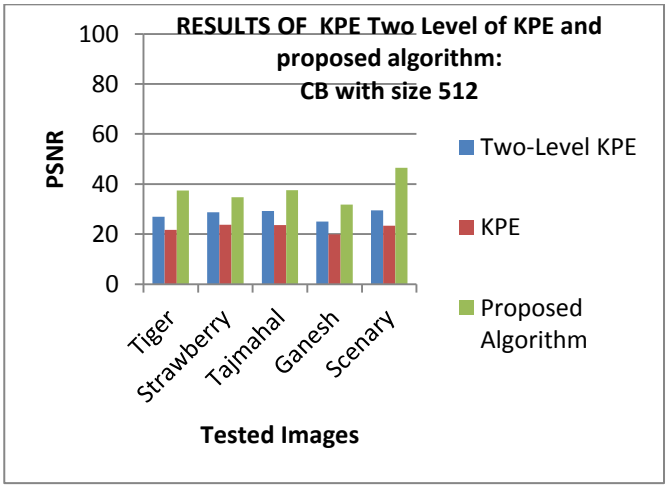


Fig. 6 PSNR of different algorithms for CB with size 512

The PSNR were calculated for different images under different algorithms shown in Table IV. In Table IV the values of the PSNR are illustrated that shows that the PSNR of proposed algorithm is high for all tested images.

ISBN: 978-1-61804-236-1

Table. IV comparison of PSNR of (a) SPHIT + SOFM based compression scheme (b) proposed algorithm (SPHIT + binary CB based compression scheme)

Test Images 512x512	SPHIT + SOFM Based Compression Scheme	Proposed Algorithm (SPHIT + Binary CB Based Compression Scheme)
Lena	32.14	39.26
Barbara	30.39	34.17
Boat	30.97	34.99
Man	31.90	36.01
Couple	30.64	35.28
Hill	30.98	35.08

Fig. 7 illustrate the change in the peak-signal-to-noise-ratio for (a) SPIHT + SOFM based compression scheme (b) Proposed Algorithm (SPHIT + Binary CB Based Compression Scheme). As shown from the chart, the PSNR remains high for all test images.

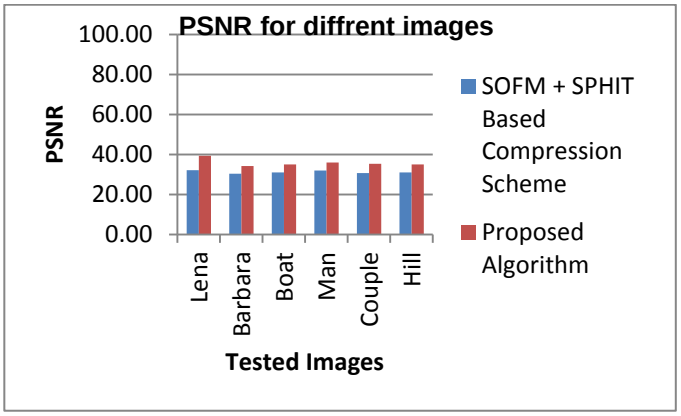


Fig. 7 comparison of PSNR of (a) SPIHT + SOFM based compression scheme (b) Proposed Algorithm (SPHIT + Binary CB Based Compression Scheme)

The CR were calculated for different images for (a) SPIHT + SOFM based compression scheme (b) Proposed Algorithm (SPHIT + Binary CB Based Compression Scheme) as illustrate in Table V. In Table V the values of the CR are illustrated that shows that the CR for proposed algorithm is less than SPIHT + SOFM based compression scheme.

Table. V comparison of CR between (a) SPIHT + SOFM based compression scheme (b) Proposed Algorithm

Test Images 512x512	SOFM + SPHIT Based Compression Scheme	Proposed Algorithm
Lena	14:01	8:01
Barbara	14:01	8:01
Boat	13:01	8:01
Man	16:01	8:01
Couple	15:01	8:01
Hill	22:01	8:01

As for Fig. 8 illustrate Comparison of CR between (a) SPIHT and SOFM based compression scheme (b) Proposed Algorithm (SPHIT + Binary CB Based Compression Scheme).

Fig.8 Comparison of CR of (a) SPIHT + SOFM based compression scheme (b) Proposed Algorithm

IV. CONCLUSION

In this paper, an improved image compression model has been introduced. The proposed scheme has been integrated with set partitioning in hierarchical trees (SPHIT) encoding to provide high PSNR. The embedded wavelet transform with SPHIT and binary codebook achieve efficient and high quality of reconstruction images comparing with other algorithms as shown in experimental results.

REFERENCES

[1] A. Karthikeyan, T. Shankar, V. Srividhya, S. Sarkar and A. Gupte,"energy efficient distributed image compression using jpeg2000 in wireless sensor networks (wsns)," Theoretical and Applied Information Technology J., vol. 47, no.3, pp. 875-883, jan. 2013.

[2] Y. Xiaobo, S. Lijuan and W. Ruchuan. (2010). Distributed Image Compression Algorithm in Wireless Multimedia Sensor Networks, http://wwwen.zte.com.cn/en/ [Online]. no. 1. Available: http://wwwen.zte.com.cn/endata/magazine/ztecommunications/2010Year/no1/articles/201003/t20100321_181536.html [Accessed 26th Oct. 2013]

[3] M. H. Yaghmaee and D. A. Adjeroh, "Priority-based rate control for service differentiation and congestion control in wireless multimedia sensor networks", Computer Networks J., vol. 53, no. 11, pp. 1798–1811, July 2009.

[4] H. Wu and A. A. Abouzeid, "Energy efficient distributed jpeg2000 image compression in multihop wireless networks," presented at the 4th Workshop on Applications and Services in Wireless Networks (ASWN), Boston, MA, August 9-11, 2004, pp. 152-160.

[5] M. Nasri, A. Helali, H. Sghaier and H. Maaref, "Adaptive image transfer for wireless sensor networks (WSNs)," in Proc. 5th International Conf. Design and Technology of Integrated Systems in Nanoscale Era (DTIS), Hammamet, 2010, pp. 1-7.

[6] M. Wu and C. W. Chen, "Multiple bitstream image transmission over wireless sensor networks," in Proc. of IEEE Sensors, vol.2, pp.727-731, Oct. 2003.

[7] H. Wu and A. A. Abouzeid," Energy efficient distributed image compression in resource-constrained multihop wireless networks," Computer Communications J., vol. 28, no. 14, pp. 1658-1668, Sept. 2005.

[8] Q. Lu, W. Luo, J. Wanga and B. Chen, "Lowcomplexity and energy efficient image compression scheme for wireless sensor networks", Computer Networks J. , vol. 52, no. 13, pp. 2594–2603, Sept. 2008.

[9] I. Andreopoulos, Y. A. Karayianris and T. Stouruaitis, "A hybrid image compression algorithm based on fractal coding and wavelet transform,"

IEEE International Symposium on Circuits and Systems, vol. 3, pp. 37-40, May 2000.

[10] S. Esakkirajan, T. Veerakumar, V. Senthil Murugan and P. Navaneethan "Image Compression using Multiwavelet and Multi-stage Vector Quantization," International Journal of Signal Processing (IJSP) J., vol. 4, no.4, pp. 246-253, 2008.

[11] M. Antonini, M. Barlaud, P. Mathieu and I. Daubechies,"Image coding using wavelet transform," IEEE Trans. Image Processing, vol. 1, pp. 205-220, Apr. 1992.

[12] H. B. Kekre and T. K. Sarode ,"New Clustering Algorithm for Vector Quantization using Rotation of Error Vector," International Journal of Computer Science and Information Security (IJCSIS) J., vol. 7, no. 3, pp. 159-165, 2010.

[13] T. W. Fry and S. A. Hauck, "SPIHT image compression on FPGAs," IEEE Trans. Circuits and Systems for Video, vol. 15, pp. 1138-1147, Sept. 2005.

[14] A. Said and W. A. Pearlman. SPIHT Image Compression: Properties of the Method. Center for Image Processing Research [Online]. Available: http://www.cipr.rpi.edu/research/SPIHT/spiht1.html[accessed 29/10/2013].

[15] D. Rawat and S. Meher,"A Hybrid Coding Scheme Combining SPIHT and SOFM Based Vector Quantization for Effectual Image Compression," European Journal of Scientific Research J., vol. 38, no. 3, pp. 425-440, 2009.

[16] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," IEEE Trans. Circuits and Systems for Video Technology, vol. 6, no. 3, pp. 243-250, June 1996.

[17] H. B. Kekre and T. K. Sarode,"Two-level Vector Quantization Method for Codebook Generation using Kekre's Proportionate Error Algorithm ,"International Journal of Image Processing J., vol. 4, no. 1, pp. 1-10 , 2010.

[18] R. D. Dony and S. Haykin, "Neural network approaches to image compression," in Proc. of the IEEE , vol. 83, no. 2, pp. 288-303, Feb. 1995.

[19] V. Dubey, N.K. Mittal and S.G. kerhalkar,"A Review on Wavelet-Based Image Compression Techniques,"International Journal of Scientific Engineering and Technology J., vol. 2, no. 8, pp. 783-788, Aug. 2013.

GPIP: A new Graphical Password based on Image Portions

Arash Habibi Lashkari
Postgraduate Center of Studies (PGC),
Limkokwing University of creative technology (LUCT), Cyberjaya, Malaysia
A_Habibi_L@hotmail.com

Abstract — Among the most significant issues in information security is the graphical password or user authentication process. Presently, alphanumeric passwords are the most widespread and well-established authentication approach in use. The technique of using the text passwords to authenticate users was introduced in the late 1960s and up-to-date most computer systems, networks and applications use this scheme. This approach has gradually become susceptible to several vulnerabilities and drawbacks due to the phenomenal increase in users and services. In order to replace the text passwords in 1996 Blonder proposed a new secure technique called graphical password which uses images and pictures instead of text. In comparison to text-based passwords, graphical authentication mechanisms guarantee improved security, memorability and usability. This paper proposes and evaluates a new graphical password scheme called GPIP for smart phones and tablets with the purpose of improving the log-in process in both security and usability aspects.

***Keywords*—** Graphical password, recognition-based schemes, pure recall-based schemes, cued recall-based schemes

I. INTRODUCTION

Information security is a major concern regardless of the size and nature of an organization. The rapid interconnectivity of businesses globally increases the importance of protecting information and the implementation of adequate security mechanisms with respect to confidentiality, integrity and authenticity. [1]. Perhaps the most widespread and convenient authentication method is the traditional textual password. This is largely to the fact that users are familiar with it and it is easy to use, and cheap to implement.

The tendency of users electing to use passwords with predictable characteristics, which in turn reduces password strength and makes it vulnerable to various attacks, is a well known weakness of traditional user authentication [2, 6]. For a password to provide adequate security it should be at least eight characters or longer, random, without any semantic content, with mix of uppercase and lowercase letters, digits, and special symbols. Even though users are aware of tips and recommendations, they typically ignore them. Some user even go as far as writting down their passwords on a piece of paper, sharing their passwords with other users or using the same password for multiple accounts [3,8]. These weaknesses perpetuated by users are the ones used as vehicles of penetrating systems in most of the common attacks such as brute force search attack, dictionary attack, guessing attack, shoulder surfing attack, spyware attack, and social engineering.

Graphical password schemes have emerged to try and overcome the weaknesses of traditional textual password and possibly enhance security. Graphical passwords are easier to remember because human beings have a natural ability to better recognize visual information as opposed to verbal information [4]. In 1996, Greg Blonder introduced the first graphical password based scheme. In his scheme a password is created by the user clicking on several locations on the image. In the same way the user must click on previously selected locations on the image or close to those locations to login [5]. Even though most of the graphical password authentication schemes have not been widely adopted, there is a growing interest in graphical passwords.

II. LITERATURE REVIEW

User authentication mechanisms are currently categorized into three main types: biometric authentication (something you are), token-based authentication (something you have), and knowledge-based authentication (something you know) [6,24].

Biometric authentication refers to the identification of some unique physical or behavioral characteristics of the user. Examples include fingerprint, iris scan, handwritten signature, voice recognition, and more. Despite the fact that biometric passwords are very efficient, easy to manage and do not require memorizing, they are expensive solutions, which cannot be widely adopted [7, 16].

Token-based authentication is a technique where in order to be authenticated the user is required to present a token. Unfortunately, the token can be easily stolen, forgotten or duplicated. Also token-based authentication scheme is not convenient for use because special additional hardware devices are needed [8, 14].

Knowledge-based authentication can be classified into two categories: textual passwords and graphical passwords. Graphical passwords include recognition-based techniques and recall-based techniques. Using **recognition-based** techniques, in order to pass the authentication, a user is required to recognize and identify a set of images selected earlier during the registration phase. Recall-based techniques are categorized into: pure recall-based and cued recall-based. In the **pure recall-based** category, the user is asked to recall and reproduce something created or selected earlier during the

Dr. Arash Habibi Lashkari has more than 7 years experience as a researcher in computer security especially in authentication processes (a_habibi_l@hotmail.com).

registration phase without being given any hint. Where as in the **cued recall-based** category, the technique proposes a hint that helps the user to recall and reproduce previously created or selected password more accurately [9,18].

1. RECOGNITION-BASED AUTHENTICATION SCHEMES

1.1 Déjà vu algorithm

In 2000 Dhamija and Perrig proposed a new graphical authentication scheme called Déjà vu algorithm, which is based on the perception of hash visualization technique. At registration phase the user is asked to choose a certain number of images from a collection of random non-describable abstract pictures generated by a system. Later, the user will be required to identify previously selected images in order to be authenticated [7]. The average registration and login time of this approach is much longer than in the traditional text-based approach. Also the server needs to store a large number of pictures that may delay the authentication process while being transferred over the network. Furthermore, the process of selecting and identifying a set of images from the picture database can be time consuming for the user [5].

Fig 1. An example of Déjà vu algorithm

1.2 Triangle algorithm

In 2002 Sobrado and Birget developed a new graphical password scheme called Triangle algorithm that is aimed to deal with the shoulder surfing problem. At registration phase a user is asked to choose a certain number of pass objects from 1000 proposed objects. Later, to authenticate, the system displays a variety of objects on the screen and the user is asked to click inside the area that they previously selected objects from. The action repeats several times but every time the icons on the screen will shuffle and appear in different locations [7]. The major disadvantage of this scheme refers to a very crowded display, so the user cannot distinguish the objects on the screen [5].

Fig 2. An example of Triangle algorithm

1.3 Passface algorithm

In 2000 Brostoff and Sasse from Real User Corporation proposed a new graphical authentication scheme that is called Passface algorithm. To create a password the user will be asked to choose a certain number of images of human faces from the picture database. At authentication phase user will be required to identify previously chosen faces in order to be authenticated. The user recognizes and clicks on the known face, and then the procedure repeats several times. The majority of the users tend to choose faces of people based on the obvious behavioral pattern, which makes this authentication scheme kind of predictable and vulnerable to various attacks [9].

Fig 3. An example of Passface algorithm

2. PURE RECALL-BASED AUTHENTICATION SCHEMES

2.1 Draw-a-Secret (DAS) algorithm

In 1999 Jermyn, Mayer, Monrose, Reiter, and Rubin proposed a new graphical password scheme called Draw-a-Secret algorithm. This scheme allows user to draw a unique password on a 2D grid. At registration phase the coordinates of the grids occupied by the drawn patterns are stored in order of the drawing. During authentication phase, the user is asked to redraw the picture by touching the same grids and in the same sequence [6]. Unfortunately, most of the users over a certain period of time forget their drawing order. Another drawback is that the users tend to choose weak graphical passwords, which as a result makes this authentication scheme kind of predictable and vulnerable to various attacks [5].

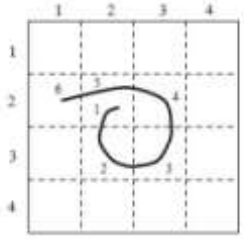


Fig 4: An example of Draw-a-Secret (DAS) algorithm

2.2 Grid selection algorithm

In 2004 Thorpe and Oorschot proposed a new graphical authentication scheme that is called Grid selection algorithm. Firstly, within a large selection grid the user chooses a smaller grid for drawing. This adds an extra degree of complexity to the password. Then the user zooms in this piece of grid and creates a drawing like in the original Draw-a-Secret (DAS) scheme. This technique of authentication dramatically increases the password space. However, it introduces additional steps i.e. to memorize and time to input the password. In other words, the security enhancement is achieved by sacrificing password usability and memorability [7].

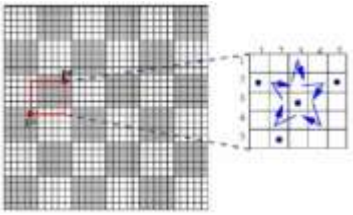


Fig 5. An example of Grid selection algorithm

2.3 Syukri et al. algorithm

In 2005 Syukri, Okamoto, and Mambo proposed a new graphical authentication scheme called Syukri et al. algorithm. During the registration phase the user will be asked to draw the signature with an input device. At verification phase the system extracts the parameters of the signature that are stored in the database. The biggest advantage of this approach is that signatures are hard to fake [16]. Also there is no extra work of memorizing the password. The main drawback is that drawing the signature with a mouse is not an easy task. The obvious solution to this problem would be the use of a pen-like input device instead of a mouse. However, such devices are not widely used and adding new hardware can be expensive [14].

Fig 6: An example of Syukri et al. algorithm

3. CUED RECALL-BASED AUTHENTICATION SCHEMES

3.1 Blonder algorithm

In 1996 Blonder proposed a new graphical authentication scheme that is called Blonder algorithm. During the registration the user is asked to click on several locations on an image to create a password. At authentication phase the user has to click on previously selected locations on the image or close to those locations. The image acts as a hint for the user to recall graphical passwords and therefore this method of authentication is considered more convenient than unassisted pure recall-based schemes [6]. The major problem this scheme is faced with, is the number of predefined click areas is relatively small so the password has to be quite long to be secure. Also, the usage of predefined click areas requires simple and plain images, instead of complex, real-world and crowded scenes [14].

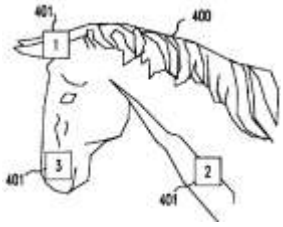


Fig 7. An example of Blonder algorithm

3.2 Passlogix v-Go algorithm

In 2002 Passlogix Inc. Company developed a new graphical authentication scheme called Passlogix v-Go algorithm. At registration phase the password is created by a chronological situation with repeating a sequence of actions. In this method the user is asked to click on various items on the image in the correct sequence in order to be authenticated [3]. One drawback is that this technique provides only a limited password space, therefore causing the password to be kind of guessable or predictable [14].

Fig 8. An example of Passlogix v-Go algorithm

3.3 PassPoint algorithm

In 2005 Wiedenbeck, Waters, Birget, Brodskiy, and Memon proposed a new graphical authentication scheme that is called PassPoint algorithm. During registration the user is asked to click on several locations on an image. At authentication phase the user has to click on previously selected locations on the image or close to those locations. This method covers the limitations of the Blonder algorithm because the images that are used for this method should be rich enough, complex and crowded. Any pixel in the image is a candidate for a click point so there are thousands of possible memorable points and combinations [6]. One drawback is that it takes more time to input the password than text-based password users spend [5].

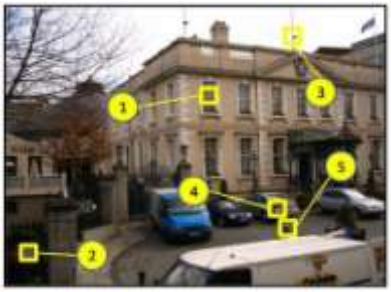


Fig 9. An example of PassPoint algorithm

III. SECURITY

A user authentication mechanism's primary goal and the main requirement is security. Likewise many strategies that exist are primarily for attacking authentication to the systems. Therefore schemes must be evaluated according to their vulnerabilities and susceptibility to different attacks since there are no system that offer perfect security.(Table 1) [11].

Brute force search attack attempts to decipher the password by searching and testing for all possible combinations of

alphanumeric characters until the correct key is found. For some graphical password schemes the most effective way against brute force search attack is to enlarge the password space by increasing the capacity of the picture library. In general, graphical passwords are less vulnerable to brute force search attacks than the traditional text-based approach. However, recall-based methods of authentication tend to have bigger password spaces than the recognition-based technique.

Dictionary attack attempts to reveal the password by running through a possible series of dictionary words that are compiled based on knowledge or assumptions considering the user's typical behavior. In general, graphical passwords are less vulnerable to dictionary attacks than the traditional text-based approach. Users of recognition-based methods usually use a mouse for input, so it has no purpose to carry out the dictionary attacks against this kind of graphical authentication. Employment of a dictionary attack for recall-based methods is much more complex than in text-based dictionary attacks but the speed of retrieving the password is slow.

Shoulder surfing attack refers to obtaining the password of a particular user during login through direct observation or using external recording devices. Text-based passwords like most of the graphical password schemes are vulnerable to shoulder surfing attacks. Only a few of recognition-based techniques are designed to resist shoulder surfing and none of the recall-based techniques are considered resistant to shoulder surfing.

Guessing attack is a very common problem for both textual and graphical authentication approaches because users usually create short and simple passwords that are convenient for guesswork. Users of the text-based approach and the users of some graphical password schemes tend to choose weak passwords with predictable characteristics.

Spyware attack refers to any unauthorized software installed without the user's permission that collects information about a user's computational behavior by tracking the keyboard input. In general, graphical passwords are less vulnerable to spyware attacks than the traditional text-based approach. Since for inputting the graphical password users exploit the mouse, the mouse motion alone is not enough to break graphical passwords.

Social engineering attack includes any method used to gain access to the system under false pretenses by exploiting human psychology. In general, graphical passwords are less vulnerable to social engineering attacks than traditional text-based approach. Also it reduces the possibility of revealing the password because the explanation of graphical password to another person by verbal interpretation is much more difficult [14, 26, 27, 28].

Table 1: Types of common attacks

Attacks	Brute force search
	Dictionary
	Guessing
	Shoulder surfing
	Spyware
	Social engineering

IV. USABILITY

In order to satisfy user needs and requirements, the development of an effective authentication mechanism should not only focus on security. The agenda should include aspects of prime importance such as usability of authentication scheme as well as its security. Included also are aspects such as learnability, efficiency of use, memorability, nice interface and overall user satisfaction with the product (Table 2) [7]. The requirement of the many steps taken by graphical authentication schemes to execute and larger amount of time spent during the registration and login phases than traditional text-based approach is very apparent. For instance, for the recall-based schemes the error tolerance has to be carefully set since the major issue is the accuracy of a user's input and its recognition during the verification phase. Further graphical authentication needs more storage space and centralized database management to store the hundreds of pictures as opposed to text-based passwords [3].

Table 2: The usability features and attributes

Usability features	Attributes
Effectiveness	Reliability and Accuracy
Efficiency	Applicable
Satisfaction	Easy to use Easy to create Easy to memorize Easy to execute Nice interface Easy to understand Pleasant picture

V. PROPOSED ALGORITHM

The objective of this paper is to propose a new algorithm for smart phones and tablets. In our proposed algorithm, users can use an image from the gallery or take a photo with the smart phone camera and add it to the algorithm as part of the original image. The system will divide the image into four sections (Figure 10). In the process of authentication user’s drug and drop the four portions which the system created each portion from one section. The system has two phases namely registration and log-in.

Figure 10: Dividing image to the four portions

Registration phase: To create a password the user chooses one image from the library or takes a photo and submits to the system. The password will be created from this image during the log-in phase (Figure 11).

Fig 11: Registration phase of GPIP algorithm

Login phase: in this part user can select four degrees of solidity by selecting 1, 2, 3, or 4 portions as login process (Figure 12). The process of user authentication is based on four sections of an image (Figure 10). Three or four portions will be created from sections and show in the list of selectable portions below four empty sections (Figure 13). The user then drugs and drop each portion in the actual section on top (figure 14). If the user drops all portions in the true sections the user can log-in to the system.

Figure 12: Four degree of solidity

Figure 13: Selectable portions for the user in login phase

Figure 14: A user tries to login

VI. Evaluation

The test and analysis method of evaluation in this project was real log-in processes by 50 users and attacks by five attackers on Android based hand phones. After developing the application using java for android phones, the application was distributed to 50 students who own smart phones running on Android operating system. Each user tries to login to the system 20 times (after two weeks working with App) and attackers sitting behind them try to figure out the password. The attackers' team comprised of 5 professional hackers from our security research center. Each attacker had worked for more than two weeks on the application and had sufficient experience on this algorithm. The results showed that the attackers' team was 16% successful in this testing process and the users were more than 92% satisfied and 90% successful in log-in process with this application. The major factor which all users mentioned was that they selected their image or used the phone camera to take their personal picture for the original picture of application.

VII. Results and Discussion

Table 1 lists the common attacks in graphical passwords and based on previous research, the results of the GPIP test and the analysis process. A comparison of the different schemes is summarized in Table 3 [26, 27, 28]. In the table just the * cells have tested in this research on the real environment and other cells founded from previous researches.

Table 3: Security Evaluation

Algorithms		Attacks					
		Brute force	Dictionary	Shoulder Surf.	Guessing	Spyware	Social eng.
Recognition	1.1 Déjà vu	👎	👍	👎	👎	👍	👍
	1.2 Triangle	👎	👍	👍	👎	👍	👍
	1.3 Passface	👎	👎	👎	👎	👍	👍
	1.4 GPIP *	👍	👍	👈	👍	👍	👍
Pure recall	2.1 DAS	👍	👎	👎	👎	👍	👍
	2.2 Grid selection	-	-	👎	-	-	-
	2.3 Syukri et al.	👍	👎	👎	👎	👍	👍
Cued recall	3.1 Blonder	👎	👍	👎	👎	👍	👍
	3.2 Passlogix v-Go	👎	-	👎	👎	-	-
	3.3 PassPoint	👎	👍	👎	👎	👍	👍

A comparison of usability attributes (Table 2) in different graphical password schemes is summarized in Table 4. In the table just the * cells have tested in this research on the real environment and other cells founded from previous researches.

ISBN: 978-1-61804-236-1

Table 4: Usability Evaluation

Algorithm		Effectiveness	Efficiency	Satisfaction						
		Reliable , Accurate	Applicable	Easy to use	Easy to create	Easy to memorize	Easy to execute	Nice interface	Easy to understand	Pleasant picture
Recognition	1.1 Déjà vu	👍	👍	-	👍	-	👍	👍	-	👍
	1.2 Triangle	👍	👎	👍	-	👍	-	-	👎	👎
	1.3 Passface	👍	👍	-	👍	👍	👎	👍	👍	👍
	1.4 GPIP *	👍	👍	👍	👍	👍	👍	👍	👍	👍
Pure recall	2.1 DAS	👍	👍	-	👎	👎	👍	-	👍	👎
	2.2 Grid selection	👍	-	👍	👎	-	👍	👎	👍	👎
	2.3 Syukri et al.	👍	👍	👍	-	👍	-	👍	👍	👎
Cued Recall	3.1 Blonder	👎	👍	-	-	👍	-	👎	👍	👎
	3.2 Passlogix	👍	👍	-	👎	👍	👎	👍	-	👎
	3.3 PassPoint	👍	👍	-	👍	👍	-	👍	👍	👍

VIII. Conclusion

This paper reviewed algorithms on recognition-based, pure recall-based and cued recall-based methods. We discussed the security and usability issues of graphical authentication including six common attacks based on previous works. From our study it is evident that most of the graphical password schemes are vulnerable to brute force search, shoulder surfing and guessing attacks. Overall both usability and security of an authentication scheme are of prime importance. This paper proposed a new graphical password namely GPIP. The algorithm developed was implemented on smart phones using java. The evaluation was undertaken by 50 users while 5 attackers attempted to attack the algothrim. The result shows that the system is more than 85% secure and more than 90% user friendly. Finally the attackers' team suggested adding some decoy portions in the selectable portions part which could improve the security of algorithm.

References

[1] Meng, Y. (2012) Designing Click-Draw Based Graphical Password Scheme for Better Authentication, *IEEE Seventh International Conference on Networking, Architecture, and Storage*

[2] Hu, W., Wu, X. & Wei, G. (2010) The Security Analysis of Graphical Passwords, *International Conference on Communications and Intelligence Information Security, China*

[3] Ma, Y. & Feng, J. (2011) Evaluating Usability of Three Authentication Methods in Web-Based Application, *Ninth International Conference on Software Engineering Research, Management and Applications, USA*

[4] Eljetlawi, A. & Ithnin, N. (2008) Graphical Password: Comprehensive study of the usability features of the Recognition Base Graphical Password methods, *Third International Conference on Convergence and Hybrid Information Technology*

[5] Lashkari, A. H., Towhidi, F., Saleh, R. & Farmand, S. (2009) A complete comparison on Pure and Cued Recall-Based Graphical User Authentication Algorithms, *Second International Conference on Computer and Electrical Engineering*

[6] Wiedenbeck, S., Waters, J., Birget, J., Brodskiy, A. & Memon, N. Authentication Using Graphical Passwords: Basic Results.

[7] Suo, X., Zhu, Y. & Owen, G.S. Graphical Passwords: A Survey

[8] Wiedenbeck, S., Waters, J., Birget, J., Brodskiy, A. & Memon, N. Authentication Using Graphical Passwords: Effects of Tolerance and Image Choice

[9] Gao, H., Liu, X., Dai, R., Wang, S. & Liu, H. (2009) Design and Analysis of a Graphical Password Scheme, *Fourth International Conference on Innovative Computing, Information and Control*

[10] Almulhem, A. A Graphical Password Authentication System, Saudi Arabia

[11] Sabzevar, A. & Stavrou, A. (2008) Universal Multi-Factor Authentication Using Graphical Passwords, *IEEE International Conference on Signal Image Technology and Internet Based Systems.*

[12] Qureshi, M., Younus, A. & Khan, A. (2009) Philosophical Survey of Passwords, *IJCSI International Journal of Computer Science Issues, Vol. 2, Pakistan*

[13] Eljetlawi, A. (2008) Study and Develop a New Graphical Password System

[14] Lashkari, A. & Towhidi, F. (2010) Graphical User Authentication (GUA), *LAP LAMBERT Academic Publishing, Germany*

[15] Tao, H. (2006) Pass-Go, a New Graphical Password Scheme

[16] [16] Biddle, R., Chiasson, S. & Oorschot, P. (2011) Graphical Passwords: Learning from the First Twelve Years

[17] Chiasson, S. (2008) Usable Authentication and Click-Based Graphical Passwords

[18] Thorpe, J. & Oorschot, P. (2004) Towards Secure Design Choices for Implementing Graphical Passwords, *20th Annual Computer Security Applications Conference, IEEE*

[19] Lasarus, R. (2006) Pass-Color: Generating Password with Colored Graphical Assistance

[20] Li, Z., Sun, Q., Lian, Y. & Giusto, D. (2005) An Association-Based Graphical Password Design Resistant To Shoulder-Surfing Attack, *IEEE*

[21] English, R. & Poet, R. (2012) The Effectiveness of Intersection Attack Countermeasures for Graphical Passwords, *IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*

[22] Dunphy, P., Heiner, A. & Asokan, N. (2010) A closer look at recognition-based graphical passwords on mobile devices, *Sixth Symposium on Usable Privacy and Security*

[23] Joshi, A., Kumar, S. & Goudar, R. (2012) A more Multifactor Secure Authentication Scheme based on graphical Authentication, *International Conference on Advances in Computing and Communications*

[24] Khandelwal, A., Singh, S. & Satnalika, N. User Authentication by Secured Graphical Password Implementation.

[25] Seng, W., Khuen, Y. & Shing, N. (2011) Enhanced Graphical Password by using Dynamic Block-style Scheme, *International Conference on Information and Intelligent Computing*, Singapore

[26] Lashkari A.H. and Farmand S. (2009) A survey on usability and security features in graphical user authentication algorithms, *International Journal of Computer Science and Network Security (IJCSNS)*, VOL.9 No.9, Singapore

[27] Masrom M., Towhidi F., Lashkari A.H. (2009) Pure and cued recall-based graphical user authentication, *Application of Information and Communication Technologies (AICT)*

[28] Lashkari A.H., Saleh R., Farmand F., Zakaria O.B. (2009) A Wide range Survey on Recall Based Graphical User Authentications Algorithms Based on ISO and Attack Patterns", *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 6, No. 3

Dr. A. H. Lashkari received his B.E. degree in Software Engineering from the IAU in IRAN (1995). He then proceeded to receive his Master of Science in computer Science and Data communication from University Malaya (UM) in Malaysia (2010). Now, he has a PhD in Computer Science (Security) from University Technology of Malaysia (UTM) (2013). He has a mixture of academia and industry experience. His research interests include Network Security, Authentication process, Graphical Passwords, Attacks and Anti-attacks.

A Real-Time Web-based Graphic Display System using Java™ LiveConnect Technology for the Laguna Verde Nuclear Power Plant

Efren Ruben Coronel Flores, Ilse Leal Aulenbacher

***Abstract*—**This paper describes the architecture of a real-time Web application, which is being developed for the new monitoring system in the Laguna Verde Nuclear Power Plant. We describe technologies and methodologies that were applied to achieve a correct implementation. We also describe the main technological challenges that were faced in order to develop a Web-based real-time application. The application was developed using Java and JavaScript, making use of LiveConnect technology to achieve interoperability between both languages. HTML5, CSS3 and SVG were used to create graphic user interface of the application.

***Keywords*—**Real-Time, Nuclear Power Plant, Web Application, Process Information.

I. Introduction

GRAPHIC displays present relevant information regarding operating conditions in power plants. This information is of vital importance to operators, because it aids in decision-making and is needed for power plant operation. Therefore, graphic displays need to be highly reliable and information must be presented clearly and concisely. In the particular case of nuclear power plants, graphic displays are often subject to a series of regulations that have to do with how information is presented to operators.

This paper describes the architecture and design of a Web application that will replace the current graphic display system in the Laguna Verde Nuclear Power Plant. One of the main requirements for this application is that it needs to be integrated to an existing real-time data acquisition system (DAS). In addition, the application must meet strict performance and reliability criteria since it is a mission-critical system that must be able to operate for extended periods of time with no interruptions. Furthermore, information presented to operators must be highly precise and reliable. Another important requirement is to ensure that the application can be supported for at least 15 years. For this reason, the application must be built using well-established languages and standards. Since the power plant is under constant renovation and maintenance, graphic displays should be easy to update in a reliable manner. Finally, processing should be distributed to client workstations in order to minimize the load in DAS servers.

Taking into consideration the application requirements stated above, we decided to develop the new graphic display system as a Web-based application. Currently, Web-based applications are becoming increasingly popular and technologies used to develop them are open and well-established; thus, software lifespan for this type of applications can be expected to be considerable. These technologies also offer a higher ease of use with respect to other programming languages. Contrary to what might be expected, we show that it is possible to develop a mission-critical, Web-based application that works in real-time.

This paper first gives a brief overview of the data acquisition system to which we will integrate our graphic display system and that will be the main data source for our graphic displays. We then present the proposed architecture, explain its modules and describe how we tackled several technological challenges to achieve a high degree of performance and reliability. Finally, we explain our graphic user interface implementation.

II. Background

The Laguna Verde Nuclear Power Plant relies on an information system known as SIIP, to monitor its processes. The system consists of servers which acquire store and process data in real time. It also consists of workstations with graphic displays, which present real-time information on key processes and systems. The SIIP system is mission-critical and is integrated with a data acquisition system (DAS) known as NSAD [1], which was developed by the Electrical Research Institute of Mexico. The NSAD system can acquire data from different sources such as RTP [2], NUMAC (Nuclear Measurement Analysis and Control) and DEHC [3] (Digital Electro-Hydraulic Control) modules, among others. This system also features software modules capable of generating composed data points, which are complex data points that calculate important parameters through specialized algorithms. Such calculations include operational limits, balance of plant, security parameters, etc. The NSAD system is also capable of generating long duration historical archives known as SCAN [4], which are used to analyze transients or important events as well as to generate tabular or graphic trend reports

ISBN: 978-1-61804-236-1

III. Architecture

Based on the requirements, we analyzed and evaluated two different possible architectures: desktop and Web-based applications. Each option has its advantages and disadvantages. For instance, desktop applications are robust and can be developed in a variety of programming languages. However, these applications may become dependent on the platform in which they were developed. Regarding the development of visual displays with complex drawings, desktop applications require higher development efforts due to the lack of an easy-to-use drawing standard. Therefore, in most cases, it is necessary to resort to third-party applications.

Web applications have their disadvantages as well, which are mostly related to communication or reliability in the context of mission-critical real-time systems. However, these applications have many advantages. For instance, Web applications are widely used and thus, extensive support in the future is expected. This is mainly because they are based on easy-to-use, open and well-established standards. Another advantage is platform independence; to deploy these applications, a Web browser with support for the required technologies is all that is needed.

Fig. 1 illustrates the proposed architecture, which highlights and defines three main modules: communication, processing and presentation. Each module makes use of different technologies. Therefore, our application requires a Web browser that supports Java, JavaScript, HTML5 [6], CSS3 [7] and SVG [8]. The latter is used for vectorial drawings in graphic displays.

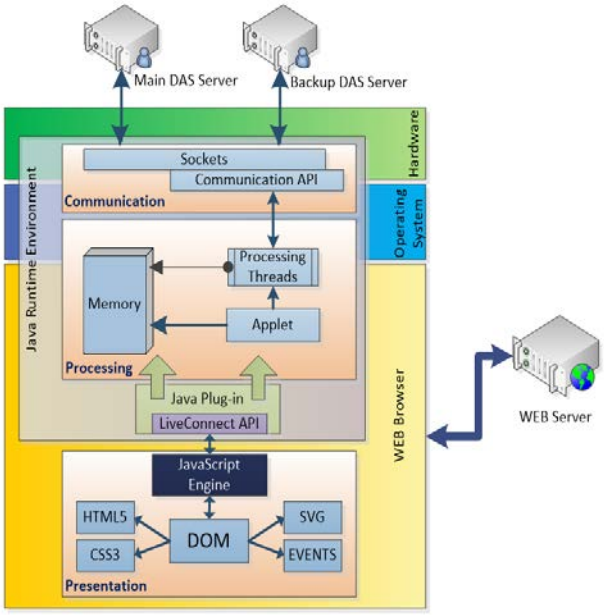


Fig. 1 General Architecture

When a Web browser loads the application, it creates a Java applet which in turn launches the Java execution environment, which is better known as the Java Virtual Machine (JVM). It is in the JVM where Java objects for both processing and communication modules are created. The JVM allows creating multiple execution threads that run asynchronously in relation to the main browser thread. From within these threads, it is possible to use sockets to establish communication with the DAS server to exchange information. Multi-threading also allows the application to perform advanced processing tasks, which require greater processing time and resources. Such complex tasks would cause performance problems if they were assigned to the main browser thread.

Finally, within the presentation module, JavaScript is used to manipulate and interact in real-time with the browser visualization environment. This is done through LiveConnect technology [9], which enables interoperability between the JavaScript engine and Java applets; which in turn, allow obtaining information stored in the JVM execution threads. LiveConnect technology is implemented in every Web browser and allows Java applets to communicate with the JavaScript engine and vice versa. Furthermore, it is in the presentation module where open standards dedicated to the visual part of the application converge: HTML5 is used to define elements such as texts, tables, labels, buttons, etc. and SVG is used to draw complex elements such as the representation of a turbine, a condenser, or a nuclear reactor vase.

IV. Communication

Communication is the main building block in real-time applications, since a high degree of performance and reliability is required. Hence, it is one of the most important requirements in our Web-based graphic display system. To meet such requirements, the main challenge was to determine which technology would be more adequate. One important consideration is that to obtain a high level of performance, the application needs to have dedicated point-to-point communication; this is achieved by using socket libraries over a well-defined protocol. Most programming languages designed for traditional desktop applications, implement sockets and their use is relatively simple. However, it is in Web environments where the use of sockets becomes more complicated.

Recently, new technologies have emerged in order to solve this problem; one of them is the Websockets standard [5], which is expected to bring real-time or close-to-real-time communication capabilities to Web environments. Websockets provide full-duplex communication over a TCP connection. There are libraries that already make use of Websockets technology, which offer developers transparent use of sockets in their applications. Nevertheless, the main disadvantage of this technology is that it is very recent and has not yet been thoroughly tested. Furthermore, it works on top of the JavaScript browser engine and thus, is part of the single thread that the browser provides. Therefore, it does not support multiple threads. This can cause serious problems, because the application could become blocked while waiting for a socket response. In a system like ours, this problem should be avoided at all costs.

ISBN: 978-1-61804-236-1

Due to the issues explained above, we decided to explore the option of using the Java language to communicate with the DAS server by using applets; this allows creating traditional Java applications that can run in a browser. With this option, it is possible to take advantage of all of the features that Java provides, which include the use of sockets and multiple threads within an execution layer in the browser. This enabled us to implement communication and data acquisition through TCP/IP sockets over Ethernet, based on the client/server paradigm. With this, we were able to guarantee the performance level we need for acquiring real-time data from the DAS server.

The communication protocol is proprietary and was designed specifically for communication with the DAS. The protocol features well-structured messages that are used to perform each of the different possible actions in the system, such as user authentication, real-time data requests, alarm information, historical data retrieval, etc. One of the advantages of having a proprietary protocol specifically designed for our system is that it offers good performance, as well as a good level of security.

In our implementation, communication sockets are declared within a Java object that can be accessed through the applet, which is instantiated in the Web Application (Fig. 2). Sockets can be used from JavaScript by using Java's LiveConnect technology, which enables interaction from JavaScript with calls to objects instantiated within the JVM by the applet itself. For example, this allows the application to perform the user authentication process by requesting a username and password to the user and then passing such parameters to the corresponding method in the Java object; this object in turn establishes the connection using sockets to send the authentication message.

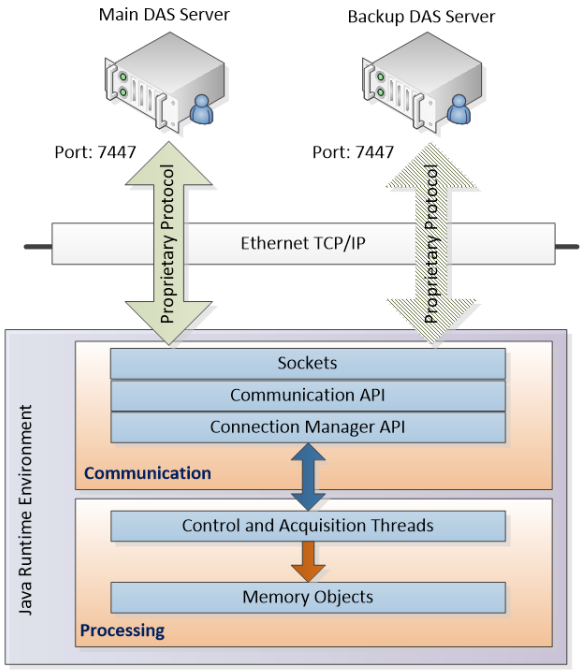


Fig. 2 Communication mechanism

The DAS can also operate in redundant mode. This means that there is a server operating as the main data server and there can be other servers working in backup mode. If the main server is down, the backup server can replace it automatically, thus avoiding any loss of information. Therefore, the communication module also has the necessary functionality to guarantee that the Web application will always be connected to one of the DAS servers. This is determined by performing periodic verifications each second, to determine which server is operating as the main data server; if the main server is down, it establishes a connection to a working server.

V. Memory Management

Memory management in real-time, Web-based applications tends to be complex, especially because most object-oriented languages contain mechanisms that dynamically create or destroy objects in memory, based on algorithms that determine when objects are no longer being used or referenced. This, in the context of real-time systems, can be problematic. Therefore, special mechanisms must be designed to properly manage objects in memory, so that application performance is not undermined. To address this problem, objects from the most important modules in our system are instantiated only once; this is done when the applet is loaded and the JVM created. In addition, these objects are static and are instantiated in a static class, which can be accessed by every object, either from Java or JavaScript. To access these objects from JavaScript, specific applet methods are invoked; these methods expose static objects stored in the JVM (Fig. 3).

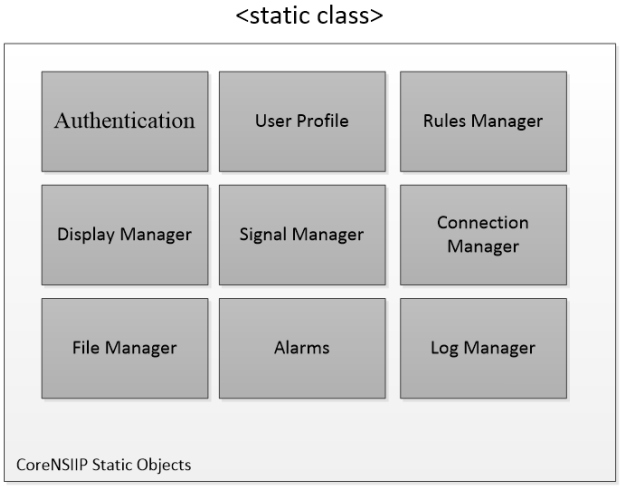


Fig. 3 Memory Objects

One issue in the interaction between JavaScript and Java is how data types are translated between both languages, when methods that return values are invoked. Native data types such as String, int, float, etc., do not cause problems. However, implementation of complex objects that include arrays can be different among browsers. To solve this problem, we use JSON objects for data exchange between JavaScript and Java. This allows us to manipulate information in a format that is common to both languages.

ISBN: 978-1-61804-236-1

Another benefit that comes from implementing correct memory management mechanisms is having greater control over Garbage Collection (GC) in Java and JavaScript. Since GC cannot be disabled, different techniques can be used to delay Garbage Collection indefinitely. One of these techniques is object pooling, which consists in re-using objects instead of constantly creating and discarding new objects. This is especially important for large or complex objects.

VI. Processing

Submission By definition, the execution of a Web application is always single threaded. This means that there is only one execution thread in which all actions, such as rendering visual components, handling events or user input, managing timers, etc., are performed. This paradigm can be problematic, especially for real-time Web applications, which have very specific requirements. One of such requirements is reliability, which means that the application must be robust enough so that it can operate without interruptions during extended periods of time. Another important requirement is performance, which enables an application to react in a timely manner under different operating conditions. This aids in avoiding deadlocks or anomalies that can result in users experiencing long waiting periods or frozen screens.

The use of Java from a Web application, allows performing several operations through multiple threads that run in parallel to the main Web browser thread, thus achieving a high level of performance and reliability. Fig. 4 shows a block diagram which illustrates the threads we implemented for our application.

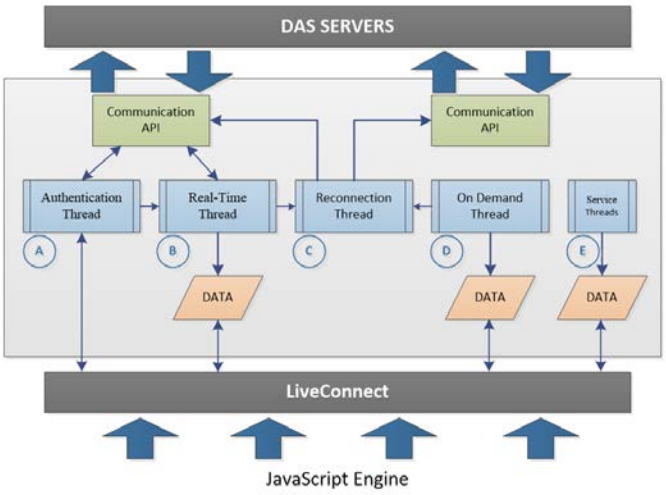


Fig. 4 Multi-Threading

A. Authentication Thread, which is created and run only once when a user first launches our application. It allows to establish a connection to the server and to send an authentication message. This thread prevents the application from becoming blocked during the authentication process. This thread also obtains certain initialization parameters, such as data point information, alarm information and server information.

B. Real-Time Acquisition Thread, which is created and run once a user has successfully authenticated and established a connection to the server. This thread operates permanently during a user session and acquires real-time data for around 8500 data points. Acquisition is performed in two cycles per second (500ms). Each cycle, data regarding software and hardware alarms, status of processes, status of acquisition subsystems and server state are obtained.

C. On-demand Data Thread, which is created and run once a user has successfully authenticated and established a connection to the server. This thread operates permanently during a user session and uses its own communication socket, which is independent from the Real-Time Acquisition Thread. The purpose of this thread is to manage user requests that involve longer processing periods, such as requests for historical data, which are often used to generate tabular or graph trend reports. These requests can take several seconds to complete due to the required data flow. If such requests were performed using the Real-Time Acquisition Thread, graphic displays would experience delays and would not be able to present the most recent information in a timely manner.

D. Reconnection Thread, which is run once, only when there is a connection problem to the server involving real-time or on-demand acquisition sockets. It also verifies that a connection is established to a server operating as the main data server. If there is no main data server online, connections are established to backup servers. Once connections have been successfully re-established, this thread exits.

E. Service Threads. These threads are created to handle requests from certain system modules that require longer processing times and that could otherwise block the execution of our Web application. For example, a request to perform an analysis on historical data for a group of data points could take several seconds or even minutes, depending on its complexity. By using a thread that is exclusively dedicated to such request, the Web browser main thread can continue to run normally, without interruption. The browser main thread would only need to monitor the corresponding service thread in order to determine when it has finished its processing, so that results can be presented to the user.

All of these threads are run asynchronously in relation to the main Web browser thread in the Java Virtual Machine, thus avoiding delays or interruptions to the browser main thread. Data which are acquired or processed by threads are stored in persistent objects in the JVM. In fact, each application module has a dedicated persistent object. These objects are accessed by JavaScript through LiveConnect calls to the Java applet, in order to obtain data which is then shown in graphic displays.

ISBN: 978-1-61804-236-1

VII. Graphic User Interface

Our Graphic User Interface (GUI) was designed as a single HTML5 application. With this approach, Web-based applications do not need to refresh a whole page as the user interacts with it, similar to traditional desktop applications. To achieve this, JavaScript provides power and flexibility by enabling developers to create or modify components by accessing and manipulating the DOM (Document Object Model) in a Web page. Additionally, AJAX technology provides the capacity to obtain dynamic data from Web servers without the need to update an entire page. Fig. 5 shows a block diagram which illustrates how our Web application works.

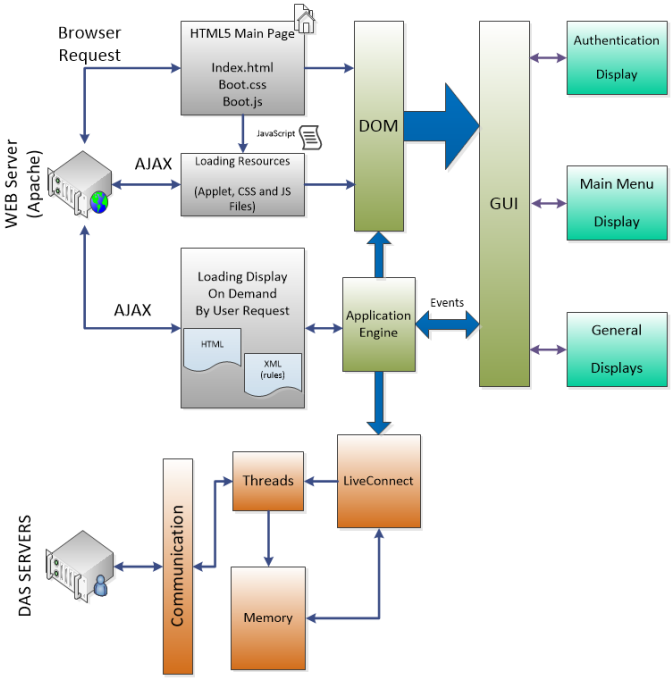


Fig. 5 User Interface block diagram

a) Our Web application consists of a single HTML index file (index.html), which includes a reference to a CSS file and a JavaScript file, which is used to load the site. Our Web application is stored in an Apache Web server, which can be accessed from any Web browser.

b) Once our Web application is loaded, a JavaScript process loads via AJAX other resources that constitute our application. These elements are: an applet which serves as an interface to Java, CSS files and JavaScript files from different modules that make up our application. The obtained resources are then inserted into the document DOM so that they become active in browser memory.

c) Once our application is loaded, an authentication screen allows users to connect to the DAS server. At this point, connection mechanisms are used to establish communication with DAS servers. It is important to note that DAS servers are different from the Web server where our Web application is hosted.

d) Once a user has authenticated successfully, the application main menu is shown. From this menu, users can open any of the graphic displays in our system. These graphic displays are implemented as HTML5 pages and can contain graphic elements based on SVG or Canvas. Graphic displays can also contain special tags which indicate whether an element has a dynamic behavior associated with it, such as showing real-time data, rendering a graph or showing information about data points or alarms. These special tags are extracted and substituted with dynamic content by the application engine, which loads the corresponding graphic display and inserts it into the DOM. It is in this post-processing phase, where graphic displays are analyzed to determine which components are to be modified dynamically.

e) The application engine is run indefinitely at a rate of four cycles per second (every 250 milliseconds). Each cycle, it obtains real-time and historical data from JVM persistent memory, for each of the data points being shown in a graphic display. These data requests are performed though LiveConnect technology by accessing applet methods, which in turn expose the requested information to the application engine via JavaScript. Additionally, the application engine dynamically modifies the content of dynamic HTML elements by manipulating the DOM and applies certain styles based on CSS files in memory. These operations are performed based on the behavior that was specified for the graphic display being shown. It is important to remember that, at this point, once a connection has been established, processing threads inside the JVM are active and acquire information from the DAS server asynchronously in relation to the main browser thread.

VIII. Conclusion

When we think about Web applications, several examples come to mind. In particular, we can think of sites that we use on a daily basis. Some of these sites are quite popular and combine some of the technologies described in this paper. However, there are very few Web-based applications that can be considered mission-critical, which operate in real-time with the level of performance and reliability required in a nuclear power plant. While designing and developing our system, we faced several major challenges, which were not easy to tackle. Despite the fact that these technologies are well-documented and widely used, it was always necessary to go one step further and discover new ways to use and integrate them. We consider that the main challenge was to achieve a correct Java applet implementation. We can say that the Java applet is the most important element in our application because it enabled us to achieve real-time communication and to implement several processing threads. The main problem with applets is also its greatest virtue: security. While the possibility of creating communication sockets, accessing disk files or accessing computer resources is very powerful, it also generates problems; therefore, it is clear that applet implementation is

not simple by any means. The use of LiveConnect technology opens a wide range of possibilities for this type of applications. Inter-communication between JavaScript and Java greatly enhances processing and performance capabilities in Web applications. Currently, our application is in the final phase of development and testing. Tests have been successful since our application complies with all requirements described in this paper.

REFERENCES

[1] I. Leal, J. Suarez, E. Coronel, "A Real-Time Data Acquisition System for the Laguna Verde Nuclear Power Plant", WSEAS, July 2010, ISSN: 1109-2750.

[2] E. Coronel, "Desarrollo de un subsistema de adquisición de datos de equipos RTP, para su uso en el nuevo sistema de adquisición de datos en tiempo real de la Central Nucleoeléctrica Laguna Verde", CIINDET, pp. 4 - 5, México, Octubre 2008.

[3] E. Coronel, C. Chairez, "Subsystem of Data Acquisition Using the ModBus Protocol in Real Time of the Digital Electro-Hydraulic Control and Its Integration with the Integral System of Process Information of Laguna Verde Nuclear Power Plant", CERMA, November 2012, pp. 153 - 156, ISBN: 978-1-4673-5096-9.

[4] I. Leal, J. Suarez, "Registros históricos de tipo SCAN en memoria para un sistema de adquisición de datos en tiempo real para la Central Nucleoeléctrica de Laguna Verde", CIINDET, México, Octubre 2008.

[5] I. Fette, A. Melnikov, "The WebSocket Protocol", IETF, December 2011, ISSN: 2070-1721.

[6] HTML5 A vocabulary and associated APIs for HTML and XHTML. World Wide Web Consortium (W3C), Editor Draft's, http://www.w3.org/html/wg/drafts/html/CR/.

[7] Håkon Wium Lie & Bert Bos: Cascading Style Sheets – designing for the Web "written by the creators of CSS" (3rd edition, Addison-Wesley, 2005, ISBN 0321193121.

[8] Scalable Vector Graphics (SVG) 1.1 (Second Edition), World Wide Web Consortium (W3C), August 2011. http://www.w3.org/TR/SVG/

[9] D. GoodMan, "JavaScript Bible 3rd Edition", Chapter 38, ISBN: 0764531883

Open sources information systems used in risk management for healthcare

Daniela Drugus, Doina Azoicai and Angela Repanovici

***Abstract*—** The paper presents the importance and benefits of open sources especially those regarding healthcare. The risk management process includes different methods of decision making considering the predictable and unpredictable risk situations. The basic principles of risk management are analyzed. Authors start from the assessment of risk management implementation need in the Romanian healthcare system. A quality marketing research based upon a previous documentation from specialized literature is performed. We used a SurveyMonkey online questionnaire and present the research results.

***Keywords*—** computerized systems, open sources, risk management, sanitary system, Romania.

I. Introduction

THE management of information within the healthcare system has developed in various directions. Medical information and documentation centers put at the disposal of the interested medical parties various possibilities for storing, archiving and accessing information. Referring to risk management, the situation is still at an individual level, there are no information instruments or procedures distributed or shared between different institutions of the healthcare system. Risk management in health care is defined "by clinical and administrative activities undertaken to identify, evaluate, and reduce the risk of injury to patients, staff, and visitors and the risk of loss to the organization itself" [1]. The 7 steps in the Risk Management process are establishing the context, "identifying, analyzing, evaluating, and treating the risks, continuous monitoring and review, and communication and consultation" [2]. The policy development and executive program in risk management system "consisted of designating a leader and coordinator core and defining its role" [3], and defining communications with hospital boards and committees, describing processes and preparing the infrastructure for patient safety education and culture-building [4].

Risk management has had reactive and proactive approaches including adverse event reporting and learning, root cause investigation and failure mode and effect analysis [5].

Daniela Drugus is with the Medicine and Pharmacy University, Iasi, (e-mail: drugus_daniela@yahoo.com).

Doina Azoicai is with the Medicine and Pharmacy University, Iasi, Romania (e-mail: doina.azoicai@gmail.com).

Angela Repanovici is with Transilvania University of Brasov, Romania, Faculty of Product Design and Environment (corresponding author , 40745820361,e-mail: arepanovici@unitbv.ro).

II. The Validated Model for Risk Management System

Health and clinical service delivery organizations are obliged to provide a safe environment for patients as well as staff [6].

Several different studies revealed that risk management is the basis for minimization of medical errors and enhancement of patient safety in hospitals which needs to be implemented as strategies and practical plans; and, simultaneously, clinical staff should be trained and well oriented of different risk management guidelines and scheme [7].

Results of different research studies have demonstrated that educating staff regarding safety measures can lead to patient safety improvement [8]. The policy development and executive program in risk management system consisted of designating a leader and coordinator core and defining its role, and defining communications with hospital boards and committees, describing processes and preparing the infrastructure for patient safety education and culture-building [5].

Risk management has had reactive and proactive approaches including adverse event reporting and learning, root cause investigation and failure mode and effect analysis. (Fig. 1).

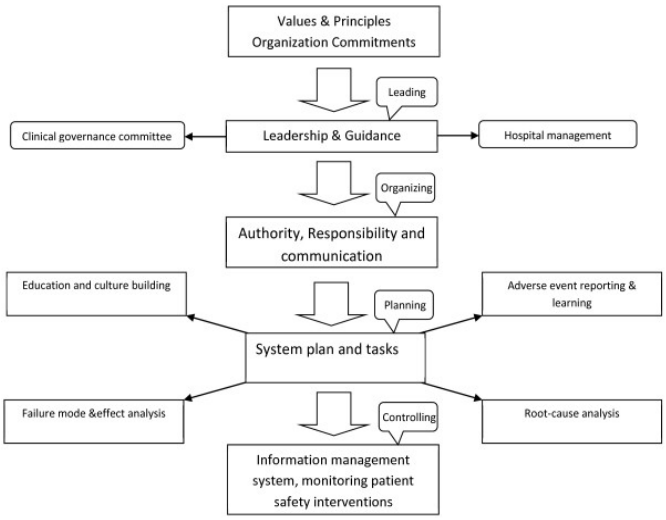


Fig. 1 The validated model for risk management system [5]

ISBN: 978-1-61804-236-1

III. Risk management in the Romanian healthcare system

The implementation of risk management systems in the healthcare sector is facilitated by the legal framework in force resulted from the transposition of European directives in the Romanian legislation and the harmonization of the national legislation with the international one. The spade work refers to training the employer as well as the employee for elaborating the risk management system. At this phase, the following actions must be taken into consideration:

- the staff attending specialized training courses,
- proper documentation on the requirements for implementing the management system,
- providing specialized consultancy for elaborating and implementing the system, consultancy for which must be contracted only widely recognized experienced institutions [9].

IV. Methodology

The authors have developed a qualitative marketing research study regarding the implementation level of the risk management system in the healthcare system. The research relied on an online questionnaire, survey monkey, https://www.surveymonkey.com/s/ML2GS72

The questionnaire was sent to the health care institutions in Iasi, Romania, a university centre with an old university tradition where there is also one of the oldest universities of medicine. The survey was conducted from January – March 2013. The survey focused on the development degree within the healthcare organizations – university, hospitals, support institutions – of the structures specialized in risk management.

V. Results

We used Survey Monkey™ to anonymously survey in health institutions.

We received answers from all the institutions which implemented risk management. The survey is validated by the gender relation, men-women, and the staff structure within the institutions taking part at this research study.

We continue with presenting the answers to the 11 questions asked.

Question no. 1 "Within your organization is there a structure supporting the process of risk management?" 97.87% of which answered "yes", 0% answered "no", and 2.13% answered "there are attempts" (Table I).

Table I Institutions supporting risk management

YES	NO	THERE ARE ATTEMPTS
97,87%	0%	2,13%

At question no. 2 "Do you have system and/or operational system procedures that regulate the activity of identification, measuring, hierarchization, treating, monitoring and documenting the risks that can affect the organization?" 95.74% answered "yes", 2.13% answered "no" and 2.13% answered "there are initiatives" (Table II).

Table II Institutions having operational system procedures

YES	NO	THERE ARE INITIATIVES
95,74%	2,13%	2,13%

At question no. 3 "Do you asses and document the risk when you take important decisions (initiating projects, drawing up strategic plans etc.)?" 95.74% answered "yes", 4.26% answered "no" (Table III).

Table III Assessment of risk initiatives

YES	NO
95,74%	4,26%

At question no. 4 "Is any type of professional training method used to facilitate and develop the amount of information referring to risks?" 95.74% answered "yes", 4.26% answered "no" (Table IV).

Table IV Professional training existence

YES	NO
95,74%	4,26%

At question no. 5 "Are there plans for emergency situations that correspond to unlikely situations, but with major consequences, which can block the organization's activity?" 80.85% answered "yes", 0% answered "no" and 19.15% answered "there are intentions of drawing them up" (Table V).

Table V Emergency plans existence

YES	NO	THERE ARE INTENTIONS OF DRAWING THEM UP
80,85%	0%	19,15%

At question no.6 "Does the organization use risk transfer or sharing instruments with other organizations (eg. Insurance companies)?" 48.89% answered "yes", 31.11% answered "no" and 20% answered "there are intentions" (Table VI)

Table VI Use of risk transfer instruments

YES	NO	THERE ARE INTENTIONS
48,89%	31,11%	20%

At question no. 7 "Is there a risk reassessment process after the implementation of the measures meant to

diminish/counteract the risk identified?" 70.21% answered "yes", 6.38% answered "no" and 23.40% answered "there are intentions" (Table VII).

Table VII Existance of risk reassesment process

YES	NO	THERE ARE INTENTIONS
70,21%	6,38%	23,40%

At question no. 8 "What are the limitations preventing you to implement the plans on diminishing risks?" lack of funds 2,12%, legislative limitations 17,02, the respondent's lack of decisional power 2,12%, absence of trained staff in this field 2,12%, the risk management information is unclear and undifferentiated on fields of activity, 2,12%, don't know 2,12%, there are no limitations 4,25% (Table VIII).

Table VIII Limitation situation in implementing preventing risks

Lack of funds	2,12%
Legislative limitations	17,02%
The respondent's lack of decisional power	2,12%
Absence of trained staff in this field	2,12%
The risk management information is unclear and undifferentiated on fields of activity	2,12%
Don't know	2,12%
The are no limitations	4,25%

At question no.9 "What do you think is the main risk for your organization?" the following major risks have been stated: insufficient financing 21,27%, reduced activity due to the decreased number of students and employees 2,12%, measures which lead to the unfulfillment of the objective 4,25%, measures taken against the organization's interests 2,12%, routine 2,12%, lack of interest in becoming familiar with the legislative field 2,12%, unknown external factors 2,12%, don't know 2,12%, there are no risks 2,12% (Table IX).

Table IX The main risk for organizations

Insufficient financing	21,27%
Reduced activity due to the decreased number of students and employees	2,12%
Measures which lead to the unfulfillment of the objective	4,25%
Measures taken against the organization's interest	2,12%
Routine	2,12%
Lack of interest with the legislative field	2,12%
Unknown external factors	2,12%
Don't know	2,12%
There are no risks	2,12%

At question no. 10 "Your gender" 80.43% have been females and 19.57% males (Table X).

Table X Gender situation

FEMALES	MALES
80,43%	19,57%

At question no.11 "Your institution" the following institutions being represented: public or state institutions, "Gr.T. Popa" University of Medicine and Pharmacy Iasi , hospital, Department of public health, Forensic Medicine Institute of Iasi (Table XI).

Table XI Institutions segmentation

Public or state institutions	44,6%
"Gr. T. Popa" University of Medicine and Pharmacy Iasi	25,53%
University	14,89%
Hospital	6,38%
Department of public health	4,25%
Forensic Medicine Institute of Iasi	2,12%

VI. Discussions

Most institutions have implemented structures for risk management consisting in a set of procedures and rules according to the legal provisions in force. Any initiative or project starts with the assessment of risks. There is a training plan in this field and an emergency situation plan, but not in all institutions. In some institutions there is only the intention of taking these measures.

Unfortunately, 31% of the respondents are not aware of any collaboration and cooperation initiatives with other institutions for reducing risks.

The majority considers risk reassessment should be conducted after the implementation of the measures, but there are institutions in which this presents itself just as an intention. The legal limitations and the insufficient founding prevent the system from developing better.

We consider that better dissemination and collaboration on the measures and good practices in diminishing risks are effective solutions and not very expensive. Using an Open source platform for dissemination, information, documentation and collaboration in the field of risk management might be the best solution, with low costs and maximum efficacy in any informational society.[10],[11],[12].

VII. Open source risk management platform

"Mirth Connect is one of healthcare integration engines, specifically designed for HL7 message integration. It provides the necessary tools for developing, testing, deploying, and monitoring interfaces. And because it's open source, there are all of the advantages of a large community of users with quality support" [13]. RMI (Risk Management Data) data were

published to Mirth Connect. Mirth Connect is an open source product created by Mirth Corporation to transform messages among formats and/or route them from one location to another. Mirth Connect creates Nationwide Health Information Network (NwHIN) in others countries, Document Submission (XDR) messages from HL7 2.x and ICD-9-CM data files [14].

The platform was developed to receive the sources of RMI in different formats, generate RMI narrative texts in structured data, normalize these data using clinical models and Consolidated Health Informatics standard terminologies [15].

VIII. Conclusions

The findings of the research study indicated that the institutions have system procedures and the culture of risk assessment exists when important decisions have to be taken. The universities include in their curriculum courses of risk management but there are no training courses, towards which people showed great interest. There are also plans for emergency situations at most institutions.

It was noticed that the obstacles preventing the implementation of risk management plans can be overcome by creating a joint source for documentation, information centers regarding the legislation in force, implementation methods and practices, the existent procedures.

Following this analysis of needs, documents and procedures included in risk management, we implemented Mirth Connect, an open source software for information management, used to integrate procedures and risk analysis. This portal will represent a collaboration instrument and it will be useful in sharing experiences.

References

[1] Joint Commission Improving America's Hospitals. (2007). *The Joint Commission's Annual Report on Quality and Safety*. Available: http://www.jointcommission.org/improving_americas_hospitals_joint_commission_annual_report_quality_safety_2012/, [accessed June 20, 2013].

[2] Tasmanian CT review team Risk management process. (2003). *Draft guidance manual for infrastructure operators*. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3598162/, [accessed June, 1, 2013].

[3] J. Miller The definition of Risk Management in Health Care. *E-How Health* [online] Available online: http://www.ehow.com/about_6619711_definition-risk-management-health-care.html, [accessed June 20, 2013].

[4] J R Coll Physicians Lond, in *PubMed Abstract*, 32(2), pp.125-129 , 1998..

[5] H. Adibi, N. Khalesi, H. Ravaghi, M. Jafari and A.R. Jeddian, Development of an effective risk management system in a teaching hospital, in *Journal of Diabetes & Metabolic Disorders*, vol. 11-15 Available: http://www.jdmdonline.com/content/11/1/15.

[6] C. Hare, C. Davies, M. Shepherd, Safer medicine administration through the use of e-learning, in *Nursing Times*, 102(16)**,** pp. 25-27, 2006.

[7] G. Neale, *Risk management in the care of medical emergencies after referral to hospital*, [online] Available http://www.ncbi.nlm.nih.gov/pubmed/9597627, [accessed June, 1, 2013].

[8] D.A. Handel, K.J. McConnell, Emergency department length of stay and predictive demographic characteristics, in *Ann Emerg Med*, 50(3), :p.70, 2007.

[9] *Stadiul actual al managementului securitatii si muncii in Romania,* cadrul legislativ, The present stage of labour and safety management in Romania, legal framework http:// www.management/managementul-muncii/Stadiul-actual-al-managementul65785.php, [accessed June 20, 2013]

[10] A. Comsa,I. Maniu,, N. Modler, W. Hufenbach, W., EC Lovasz,, V. Ciupe, An Overview of Library Automation in *Mechanisms, mechanical transmissions and robotics Book Series: Applied Mechanics and Materials,* (162) 583-588 , 2012

[11] E.C.Lovasz, D. Perju, KH Modler, A.E. Lovasz, I. Maniu, C. Gruescu, D.Margineanu,V. Ciupe, ,A. Comsa, Demonstrative Digital Mechanisms Library in Mechanisms, *Mechanical transmissions and robotics Book Series: Applied Mechanics and Materials* (162), 37-46, 2012

[12] V. Ciupe, EC Lovasz, CM Gruescu, High Quality Document Digitization Equipment, in *Mechanisms, mechanical transmissions and robotics Book Series: Applied Mechanics and Materials* (162), 589-596 , 2012

[13] Mirth Connect, Mirth Corporation (2011), http://www.mirthcorp.com/community/mirth-connect, [accessed June, 1, 2013].

[14] S. Rea, J. Pathak, G. Savova, T.A. Oniki, L. Westberg, C.E. Beebe, C. Tao, C.g. Parker, P.J. Haug, S.M. Huff and C.G. Chute, Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data. The SHARPn project, in *Journal of Biomedical Informatics*, 45, pp.763–771, 2012..

[15] Meta health Care, Mirth solutions (2013), http://www.metahealthcare.com/solutions/mirth/, [accessed June, 1, 2013]

Integrating information retrieval and static analysis to assess relationships between components and features in software systems

Dowming Yeh, Chia- Hsiang Yeh, Wei-Chen Liu, Mei-Fang Chen, and Pei-Ying Tseng

***Abstract*—**Software development is the process of constructing a software system or the software part of a system based on user requirements. The maintenance work follows the completion of the software development. An important task in software maintenance is to identify which specific component modules correspond to the implementation of a particular feature in order to gain further insight into the design of these components. Uncovering features and components module correspondence relies on the integrity of the system documents. However, the quality of most system documents degrades through time when they are not updated with the modification of code. The need to obtain a more accurate system documents can only be met by reverse engineering the source code. It is not easy for maintainers to read directly from code since it usually involves contains substantial lines of code. To uncovering features and components correspondence, the information retrieval techniques can be applied to source code, and maintainers may obtain the information on which components are most relevant to some keywords conveying the major concept of a particular feature. However, differences between program code and general text can jeopardize the accuracy of such analysis. Some characteristics in code such as the relationship among identifiers and the repetition construct should be considered. We propose a method that transforms code by dataflow analysis and loop analysis before performing information retrieval to improve precision over analyzing the original code. Maintainers can then identify more accurately which components may be impacted when certain features of a system need improving work.

***Keywords*—**information retrieval, reverse engineering, software maintenance, static analysis.

I. Introduction

With the widespread application of computer-based systems, needs of features in these systems also grow rapidly. These requirements are mostly satisfied by software. Requirements evolve as the environment and the users of a system changes. As a result, a system must be revised to incorporate update-to-date requirements. For the maintenance staff, a basic work would be to identify which specific components or modules correspond to the implementation of a particular functional feature may be impacted. This work is usually accomplished by consulting system documents. For example, in UML-based document, one can look up the related sequence diagrams to identify the possibly impacted components. Such documents may be wanting or may not be up-to-date to reflect the true design of the present version.

The quality of most system documents degrades through time when they are not updated with the modification of code. The most up-to-date system documents can be extracted from source code through reverse engineering techniques. Reverse engineering is the process of recovering design concepts or even requirements from the implementation of a system, for example, source code. It is therefore possible to produce class diagrams from code to help a software engineer understanding the structure of an existing system. To reproduce a sequence diagrams is more challenging since sequence diagrams describe dynamic behaviors of a system. One would need to add instrumental code to the original code, run the system multiple times to produce execution traces, and combine these traces to form sequence diagrams [1][2]. The dynamic analysis cost seems not justifiable, and a static analysis cannot produce a precise description [3].

In addition to source code, there are some implementation-level system documents may be utilized to help us uncovering the relationship between components and features. Such documents include descriptions in a source control system or a bug tracking system. They are up to date with the source code since these descriptions are entered along with the updated code. Some researches show that these historical descriptions can help understand the system as well as forecast errors or required maintenance resources. Combining the information in these documents with the information obtained by reverse engineering would be more useful in understanding a software system [4].

Hassan and Holt combine the information retrieved in the source control system with the reversed static dependency graph, and apply a software architecture recovery method to understand the architecture of the NetBSD open source system [5]. Our earlier research utilize the Bugzilla bug tracking system of an open source system as a source of information, and construct a corresponding relationship between software

This work was supported in part by the R.O.C. National Science Council under Grant NSC100-2221-E-017-013.

Dowming Yeh is with the Software Engineering Department, National Kaohsiung Normal University, Kaohsiung, Taiwan 824, R.O.C. (phone: 8867-717-2930; fax: 8867-605-1006; e-mail: dmyeh@nknu.edu.tw).

Chia- Hsiang Yeh is with the Computer Science and Information Engineering Department, National Taiwan University of Science and Technology, Taipei, Taiwan 106, R.O.C. (e-mail: asuna900717@gmail.com).

features and components [6]. The obtained results by this analysis cannot show the different strength of the correlation between a software feature and different components. Such information, however, can be revealed by means of text mining or information retrieval techniques.

Text mining is a process of editing, organizing and analyzing a large number of documents to provide specific users such as decision-makers or analysts with specific information such as certain characteristics of these documents associated with some abstract or keywords. In short, text mining can extract important information from the document the information which users are interested. Information retrieval is a commonly used technology for text mining. Information retrieval can effectively extract or index vocabularies from documents, and perform a full-text search, an automatic classification or summarization on documents.

When applying information retrieval techniques, the relevance of a term with a source code file usually depends on the semantic distances between terms and the frequency of occurrences of the term in the code without differentiating code from general documents [7]. However, program text should not be treated as a conventional text. A loop structure could result in a significant repetitive execution of a segment of code so that the frequency of the occurrences of a term inside a loop would be drastically different from that outside a loop. Therefore, program text should be transformed to account for looping before the analysis.

Moreover, terms in a source code are in fact identifier and most identifiers are variable names. Naming variables randomly or with abbreviations would make such variables fail to be considered as occurrences of a term under examination even with the aid of a synonym dictionary. We propose that this problem may be mitigated by taking advantage of data flow analysis. That is, semantic distances between terms should be determined not only by the lexical similarity or a synonym dictionary, but also whether two terms (variables) are connected through data flow. A more accurate calculation of the relevance of a term with a source code file would result by combining loop and data flow analysis with the information retrieval techniques.

II. Concept Location

An important task in software maintenance is to identify which specific component modules correspond to the implementation of a particular feature or concept so that software engineers could focus their attention on the design of these components before implementing changes [8][9]. After locating the source code pertaining to a concept, engineers often apply a variety of techniques and skills, such as speedy code browsing, tracing linkages in call graphs, analyzing dynamic execution results, or using tools like grep to conduct searches over some files. These techniques may be classified in three categories: search-based (for speedy browsing and grep), program structure navigation (for linking call graphs and class hierarchies), and dynamic analysis.

A. Vocabulary-Based Search

There are two kinds of search techniques for concept location in current literatures, one based on vocabulary and the other based on information retrieval. The quality of a search technique is assessed on precision and recall. The precision measure is the quantity of correct items discovered by the search over the quantity of all items discovered; while the recall measure reports the quantity of correct items discovered over the quantity of all correct items in the system. A detailed discussion of the two search techniques follows.

The most common search method for vocabulary in source code is expressing the vocabulary of interest with regular expressions so that search tools like grep may be deployed to automate the search process. However, the problem with regular expressions is the lack of flexibility in its formal representation results a low recall rate [10]. To improve the recall measure, engineers often relax the original expression to a more general representation, but this, on the other hand, produces search results of large quantity, thus reducing precision rates significantly. Moreover, these methods cannot sort results based on the relevance between the expression and a specific hit.

The low recall performance for regular expression search may be attributed to many characteristics in natural languages, such as tenses, synonyms. For example, one may use a regular expression “find” to conduct a search, but the actual implementation adopts words in different tenses, such as “found”. Such occurrences cannot match the “find” expression, failing therefore to show up in search result. For synonyms, a search based on a key word “remove” also cannot match concepts expressed as its synonym “delete” in the implementation. Line changes may also undermine searches for concepts involving two adjacent words like “find node”, if the line change happens to be inserted between these words.

To improve the low recall rate, software engineers need to broaden their search. However, rephrasing even to a slightly more general regular expression would generally result in a significant increase of search results, which effects low precision. For example, an engineer starts with the search term “printout”, and does not receive any result. To broaden his search, he then uses the term “print” instead. This would result in a large number of false hits, since “print” appears in the name of the standard output function for almost any language, such as System.out.println() in Java or printf in C. To accommodating the aforementioned line changes by rephrasing the regular expression to “find*node” would also result in many false hits like “find the term in the node.” It is too cumbersome to browse through the overly broad search results from a low precision search.

B. Information Retrieval

Information retrieval is the science of searching for

information. In the context of text document, the technology determines the similarity between a document and a query can be measured by examining the occurring frequency of the query terms in the document. A query may involve multiple words. Since a similarity coefficient is calculated, an engineer can rank various documents based on their corresponding scores [11]. This is an advantage over the regular expression search which cannot rank their search results.

Information retrieval, however, cannot avert other problems with the vocabulary search. Characteristics in natural languages such as tense still pose the reduction of recall rates for information retrieval. Information retrieval fails to consider the structure of a sentence, which also results in false hits. For example, when searching for the concept "play music," an irrelevant text like "The video should play while the music is silent" would be considered a candidate since query terms "play" and "music" both appear in the text. Conventional information retrieval does not cover synonyms, but there are some information retrieval tools allowing users to provide synonym dictionaries or even trying to infer synonyms through their frequencies and contexts [12].

In spite of the aforementioned weaknesses, researchers devise successful applications of information retrieval to identify concepts in source code, and rebuild the traceability between source code and documentation [12]. Marcus et al. apply Latent Semantic Indexing method to locate concepts in NCSA Mosaic and compare results with other approaches [13]. Cleary et al. adopt and expand a language modeling method, hyperspace analogue to language, to the concept location problem [7]. Besides source code, they leverage information from system documentations. An experimental result with comparison with other approaches is also presented.

A probabilistic information retrieval model is exploited by Canfora and Cerulo to compare a new change request and past requests. If some similarity is identified, the historical data regarding source code changes associated with past requests would suggest the impact of the current request. Change history data come from the versioned control and bug-tracked system, i.e., CVS and Bugzilla. Change impact analyses to three systems are conducted to evaluate the proposed approach [14] [15]. These researches show that there is rich information inside identifiers and comments in source code, and further improvements are possible by combining natural language processing techniques [16].

C. Program Structure Navigation

There are considerable progresses in locating concepts in source code with program structure navigation such as call graph or class inheritance [17]. These researches propose that engineers should make use of program structure to trace related source to certain concepts [18]. Hill et al. implement a tool called Dora which combines structural and lexical information. Structural information such as system call graph is automatically produced, and lexical information representing the interested feature concept is used to prune irrelevant structure edges from the complete call graph. The relevance measure is determined by term frequency by its inverse document frequency. They claim that by focusing on the relevant part of the system, their results show that such combined approach is significantly more effective than a powerful structural program analysis technique [1].

When some relevant modules or classes have been discovered, program structure can provide important path to other parts of the system that may be relevant to a concept. The effect of program structure navigation is limited, however, in the initial locating work for a new concept since related modules for a concept may not be structurally connected [19]. It is best applied to strengthen or complement search results from other method such as information retrieval, instead of being employed as the major method in concept location.

D. Dynamic Analysis And Other Methods

Software reconnaissance method infers modules implementing a feature by analyzing information from program execution [20]. Most dynamic analysis methods follow the same approach by executing a specific feature through corresponding test case and comparing its execution traces with other traces produced by other test case to locate related modules. Some of these test cases are hard to establish in the first place, particular those whose corresponding feature cannot be triggered directly by users. Poshyvanyk et al. combine Latent Semantic Indexing method with scenario-based probabilistic ranking to achieve a better result than pure Latent Semantic Indexing method. Since the scenario-based analysis is conducted dynamically, the improvement is quite costly [21].

Some researchers apply natural language processing technology to find certain aspects or concepts in requirement documents. Baniassad et al. propose the Theme method which uncovers semi-automatically aspects from requirements with natural language processing technology [22]. Find-Concept by Shepherd et al. is a semi-automatic tool based on an integrated software development environment. With its natural language processing capability, the user can improve an initial query by expanding the query with words recommended by the tool. Two case studies are given to demonstrate the improved precision and recall in identifying the relevant source code. [23]. Thorsten and De Volder respectively propose techniques that visualize historic traces of program exploration when a user searches for specific features [1][24]. ◦

From the above discussion, the search-based methods are considered to be most cost effective as the basic approach since they can find multiple locations in code related to a certain concept without much effort. Whereas, the other two methods, program structure navigation and dynamic analysis, would consume engineers much effort to bear results

III. 3. Static Analysis

In order to enhance the recall rate of information retrieval analysis, some of the characteristics of code, such as relationships between identifiers and repetition structure, should be exploited. Meanwhile, in order to avoid excessive efforts, we apply static analysis techniques, dataflow analysis and loop analysis, to extract these characteristics.

A. Dataflow Analysis

In calculating the semantic distance between query terms and terms in the source code (mostly, identifiers), the concept of data flow relationship should be incorporated besides lexical similarity and synonyms. If the lexical similarity between query terms an identifier in the source code, say X, is deemed too week, general information retrieval methods would not consider X as a relevant term. However, if the identifier X is related with another identifier Y through data flow (possibly with assignments), and the identifier Y is relevant to query terms, then the variable X should also be considered as relevant to query terms. If the dataflow relationship is taken into consideration in the information retrieval process, the occurrences of the identifier X would then contribute to the total occurrences of relevant terms in the source code, thus increasing the recall rate of the source file.

Dataflow analysis is a static analysis technique. It collects and infers the semantics of data processed by a program. A major part of the analysis involves determining definitions and usages of the identifiers in the program code, which is usually called define-use analysis. Take x = y + z as an example. The new definition of the value of x uses the value of y and z, therefore there is a dataflow relationship between x and y as well as z, but the dataflow relationship between y and z is unclear from this statement alone.

The dataflow relationship is directional, data flowing from usage to definition. To be precise, the previous example should be expressed as there is a dataflow from y to x. Another important characteristic of data flow relationship is the transitivity. That is, if there is a dataflow from y to x, and there is a dataflow from x to w, then there is a dataflow relationship from y to w. Therefore, a set of dataflow or define-use chains is formed from the identifiers in the program.

B. Information Retrieval

Static timing analysis is commonly used in the time prediction of real-time (real-time) system. To get an accurate prediction, one of the most important factors is to determine the range of the possible number of iterations for a loop, that is, the maximum and minimum numbers of execution times. The range can be calculated by engineers based on their understanding of the program. The correctness of these numbers, however, is largely influenced by the experience and insights of the engineers. A more objective and reliable approach to obtain these data is through static analysis tools.

Although it is impossible to obtain the exact value of the number of loop executions using static analysis, the upper or lower bounds on the number of times of loop executions can be obtained in most cases. Take for an example, a simple loop in Java as follows:

```
for (index = 0; index < expr; index++) body
```

Suppose the code in expr and body does not change the value of the index variable, then the number of times for the loop iteration depends entirely on the upper and lower limits of the values of expr. It is not difficult to seek the upper and lower limits of expr using static analysis [25]. For more complex loops, such as those with break commands which terminate loop prematurely or change the loop control variable inside the loop, the upper and lower limits of expr may not uniquely determine the range of the number of loop iterations. It is, however, still possible to obtain upper and lower limits by considering the conditions of break commands and statements that change the control variable in static analysis.

Healy et al describe a method consisting of three complementary parts to analyze the number of executions of a loop [26]. The first algorithm statically calculates, if possible, values of the minimum and maximum execution times of the loop (possibly with multiple exits). Secondly, when the number of repetitions depends on variables whose value cannot be determined statically, the programmer is asked to provide the possible boundaries of these variables. It is easier to estimate the value of variables than the number of times for loop iterations. The estimation of values of variable is also more reliable, although such estimation is feasible only when the programmer have a fair understanding of the program. Finally, for nested loops with the inner loop iterations depending on control variables of the outer loop, they propose to from a summation equation for iteration times, the solution to the equation is the number of times the loop is executed.

Since our main purpose is to explore the effect of loop constructs on the similarity coefficient obtained by information retrieval instead of a precise analysis of execution time, only their first algorithm is adopted in our analysis to simplify the implementation. In fact, the second part of their approach requires user with a certain amount of knowledge in a system, which is almost not possible for the intended users of our method who is not even familiar with the concepts in the system. The third algorithm, although it can determine the number of times a loop is executed more accurately, mainly results in reducing the predicted execution times by some fractional number. Compared with the effect of the first part which account for the iteration times by a factor of ten or even hundred, we think the impact of omitting the third part is acceptable for now. If the similarity coefficient is found to be greatly influenced by the loop construct, this third part should then be considered for further research.

IV. Prototype Design and Implementation

The main architecture of our prototype system is shown in Fig. 1. The prototype consists of three major modules which correspond to information retrieval, data flow analysis, and loop analysis. The loop analysis is performed first to estimate the number of iterations for each loop in a program. The program code is expanded by repeating the code of the loop body for each loop with its estimated number from the loop analysis. The dataflow analysis is then performed on the expanded code to establish the define-use chains among the identifiers. Finally, the information retrieval process starts, and the original query terms are augmented to include the related identifiers through prior established dataflow relationships.

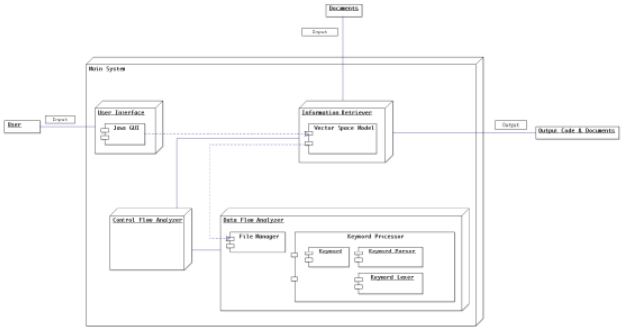


Figure 1 Prototype architecture

The activity diagram in Fig. 2 illustrates the steps in the usage of the prototype:

1. Input keywords and select source code files to be analyzed.
2. Determine whether the input is of correct form. If erroneous, an error message is generated.
3. Expand the source code from the loop analysis results.
4. Appends the addition terms from the dataflow analysis to the original keywords.
5. Performs information retrieval and record the data returned.
6. Display the analysis results.

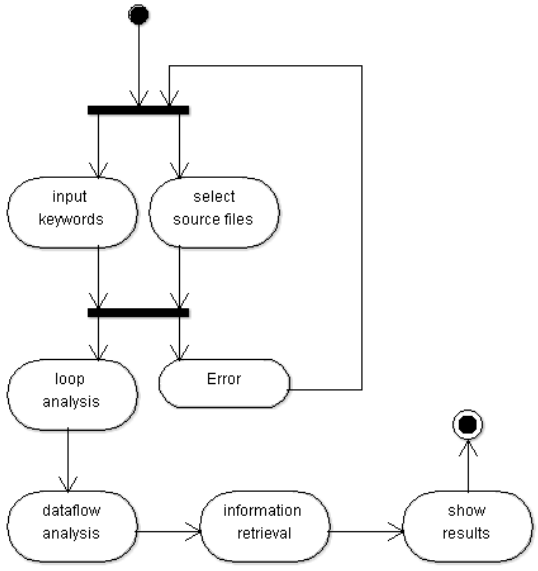


Figure 2 Activity diagram

The information retrieval module implements an algorithm based on a probabilistic model, namely, Poisson model. By computing the number of terms, the term frequency of terms, and the length of the documents, a similarity coefficient is calculated from a formula with some tuning parameters. As with other methods, a source code file is more relevant to the query terms, the similarity coefficient of a more relevant document would be larger than that of a less relevant document.

The implementation of data flow analysis is closely related to the syntactic form of programming languages. Our current implementation is targeted at Java. Since inter-procedure dataflow analysis is very complex, our prototype only performs dataflow analysis inside a procedure, i.e., a Java method. We do not need to build complete define-use chains for every program, only those that are related to the query terms are established. The related identifiers are then added to the set of query terms.

On the implementation of loop analysis, We implements the first part of the method proposed by Healy et al in the following steps.

1. Identify branches that may affect the iteration of loops
2. Determine when the direction of the loop will be changed for each branch
3. Determine the possible loop iteration numbers for each branch
4. Determine the maximum iteration number for each loop

We omit a step in the original algorithm which analyzes the equality test in the conditions that controls the branch since it will only increase the number of loop iteration by one. Also, the original algorithm can also determine the minimum iteration numbers for each loop. Yet we need to expand the code of every loop body with a specific number of times, so the maximum number is chosen to magnify the possible effect of loop on the results of information retrieval.

V. Experiments Results

To evaluate the effectiveness of various combinations of analysis approaches, we compare the computed similarity coefficients on five programs from the information retrieval (IR) technique, IR augmented with dataflow analysis (IR+D), IR with loop analysis (IR+L), and IR with both dataflow and loop analysis (IR+D+L). The results is shown in Table I. The results of IR augmented with dataflow and loop analysis are clearly improved, compared with the original IR method.

Table I. Analysis results

Program /Method	Bingo.java	Rect_oval_random.java	DrawArc2.java	ShapesTest.java	client.java
IR	0.02932	0.32877	0.06742	0.5	0.02892
IR+D	0.04398	0.61644	0.10112	0.54167	0.03614
IR+L	0.04174	0.47368	0.01644	0.66339	0.02892
IR+D+L	0.0626	0.63158	0.01722	0.66421	0.03614

ISBN: 978-1-61804-236-1

With the augmentation of dataflow analysis, some of the terms deemed irrelevant by the original IR method are considered relevant through dataflow relationship. This increases the frequency of occurrences of searched term in the code. As a result, the similarity coefficients in the row IR+D in Table 1 are all higher than those in the row IR.

The impact of loop analysis is not as evident as the dataflow analysis. If the searched terms are inside a loop, the frequency of occurrences of searched term would increase as a result of the loop expansion, which is the case for Bingo.java, Rect_oval_random.java, and ShapesTest.java. On the other hand, if the searched terms are not in any loop, the text expansion of the loop would increase the total number of terms without increasing the relevant terms, therefore reducing the value of the similarity coefficient, as in DrawArc2.java. For those programs without any loop, the loop analysis makes no difference at all, as in client.java.

VI. Conclusion

In this study, we propose a method to determine the relevance of a source code with certain features/concepts by integrating information retrieval with dataflow analysis and loop analysis. Past researches have shown that various information retrieval techniques can be applied to compute such information. However, these works treat program code as ordinary documents, which may reduce accuracies of their analyses.

We propose to analyze program code from the perspectives of data flow and control flow, so that the true information content of a program can be revealed before submitting it to an information retrieval analysis. Preliminary results show that our method renders higher similarity coefficients by applying dataflow analysis. The loop analysis also influences similarity coefficients, depending on whether the searched terms are inside the loop body or not. Experiments of larger scale are necessary to show more conclusively whether our method performs better than pure information retrieval techniques.

References

[1] T. Eisenbarth, R. Koschke, D. Simon, "Locating Features in Source Code," *IEEE Transactions on Software Engineering*, vol. 29, no. 3, pp. 210-224, Mar., 2003.

[2] E. Hill, L. Pollock, K. Vijay-Shanker, "Exploring the neighborhood with Dora to expedite software maintenance." In *Proc. 22nd IEEE/ACM International Conference on Automated Software Engineering* (ASE'07), 2007, New York, NY, USA, pp. 14-23

[3] T. Ziadi, M. A. A. da Silva, L. M. Hillah, M. Ziane, "A Fully Dynamic Approach to the Reverse Engineering of UML Sequence Diagrams," in *Proc. 16th IEEE International Conference on Engineering of Complex Computer Systems*, 2011, pp.107-116.

[4] A. E. Hassan, "Mining software repositories to assist developers and support managers," In *Proc. 22nd IEEE International Conference on Software Maintenance*, 2006, pp. 339-342.

[5] A. E. Hassan and R. C. Holt, "Predicting change propagation in software systems," *In Proc. 20th IEEE International Conference on Software Maintenance*, 2004, pp. 284-293..

[6] P. Huang, D. Yeh, and W. Lee, "Using Version Control System to Construct Ownership Architecture Documentations," in *Proc. International Conference on Data Engineering and Internet Technology*, Bali, Indonesia, Mar, 2011, pp 41-46.

[7] B. Cleary, C. Exton, J. Buckley, and M. English, "An empirical analysis of information retrieval based concept location techniques in software comprehension," *Empirical Software Engineering*, 14, pp. 93–130, 2009.

[8] T. J. Biggerstaff, B. G. Mitbander, and D. Webster, "The concept assignment problem in program understanding," In *Proceedings of the 15th international conference on Software Engineering*, 1993 , pp. 482-498.

[9] A. Marcus, V. Rajlich, J. Buchta, M. Petrenko, and A. Sergeyev, "Static techniques for concept location in object-oriented code," In *Proc. Of Int. Workshop on Program Comprehension*, 2005, pp.33-42

[10] W. Zhao, L. Zhang, Y. Liu, J. Sun, and F. Yang, "SNIAFL: Towards a static non-interactive approach to feature location," In *Porc. Int. Conf. on Software Engineering*, 2004, pp. 293-303.

[11] D. Poshyvanyk, M. Petrenko, A. Marcus, X. Xie, and D. Liu. "Source Code Exploration with Google," in *Proc. 22nd IEEE International Conference on Software Maintenance*, 2006, pp.334-338.

[12] A. Marcus, A. Sergeyev, V. Rajlich, J. I. Maletic, "An information retrieval approach to concept location in source code,", in *Proceedings. 11th Working Conference on Reverse Engineering*, Nov. 2004, pp.214-223.

[13] A. Marcus and J. I. Maletic. "Recovering documentation-to-source code traceability links using latent semantic indexing," In *Proc. of Int. Conf. on Software Engineering*, 2003, pp.125-135.

[14] G. Canfora and L. Cerulo, "Impact analysis by mining software and change request repositories." In *Proc. 11th IEEE International Symposium on Software Metrics*, 2005, pp. 9-29.

[15] G. Canfora and L. Cerulo, "Fine grained indexing of software repositories to support impact analysis." In *Proc. International Workshop on Mining Software Repositories*, 2006, pp. 105-111.

[16] A. Chen, E. Chou, J. Wong, A. Y. Yao, Q. Zhang, S. Zhang, and A. Michail. "CVSSearch: Searching through source code using CVS comments." In *Proceedings of the 17th International Conference on Software Maintenance*, Florence, Italy, 2001, pp. 364–374.

[17] M. P. Robillard and G. C. Murphy, "Concern graphs: Finding and describing concerns using structural program dependencies," In *Proc. Int. Conf. on Softw. Eng*., 2002, pp. 406-416.

[18] M. Robillard. "Automatic generation of suggestions for program investigation," in *Proc. 13th ACM SIGSOFT international symposium on Foundations of software engineering*, 2005, pp. 11-20.

[19] D. Shepherd, T. Tourwe, and L. Pollock, "Using language clues to discover crosscutting concerns," In *Proceedings of the 2005 workshop on Modeling and analysis of concerns in software*, 2005, pp. 1-6.

[20] K. Lukoit, N. Wilde, S. Stowell, and T. Hennessey. "Tracegraph: Immediate visual location of software features," in *Proc. Int. Conf. on Software Maintenance*, 2000, pp.33,39

[21] D. Poshyvanyk, Y.-G. Gueheneuc, A. Marcus, G. Antoniol, and V. Rajlich, "Feature Location Using Probabilistic Ranking of Methods Based on Execution Scenarios and Information Retrieval," *IEEE Transactions on Software Engineering*, vol.33, no.6, pp.420-432, June 2007.

[22] E. Baniassad and S. Clarke. "Theme: An approach for aspect-oriented analysis and design," in *Proc. Int. Conf. on Softw. Eng*., 2004, pp. 158-167.

[23] D. Shepherd, Z. P. Fry, E. Hill, L. Pollock, and K. Vijay-Shanker. "Using natural language program analysis to locate and understand action oriented concerns." In *Proc. International Conference on Aspect Oriented Software Development*, 2007, pp. 212-224.

[24] K. D. Volder and D. Janzen. "Navigating and querying code without getting lost," in *Proc. International Conference on Aspect Oriented Software Development*, 2003, pp. 178-187.

[25] M. E. Benitez and J. W. Davidson. "A portable global optimizer and linker," in *Proceedings of the ACM SIGPLAN conference on Programming Language design and Implementation*, New York, NY, USA, 1988, pp. 329-338.

[26] C. Healy, M. Sjödin, V. Rustagi, D. Whalley, and R. Van Engelen, "Supporting timing analysis by automatic bounding of loop iterations.," *Real-Time Systems*, 18(2-3), pp. 129-156.

A 3D Visualization of the Baťa Company's Factory Premises in Zlín in 1938

P. Pokorný and M. Vondráková

***Abstract*—**This paper describes the creation of a 3D model of the Bata factory area as it was in 1938. This year was selected due the existence of a large amount of historical materials. The first step was to create the model of Zlín's landscape and its surrounding area. For this, we used the height maps obtained from the NASA expedition. The next step was to acquire contemporary historical material, especially photos and maps. We created 3D models of the factory buildings based on these materials. For these models, we designed and drew appropriate textures. In a similar way, sub-models, such as trees or railroads were also created. After the end of the actual modeling process, these sub-models were inserted into the final scene which rendered the images and the final animation. The Blender software and Gimp were used to model and render in order to create textures.

***Keywords*—** Visualization, 3D graphics, Modeling, Texturing, Rendering.

I. Introduction

THE first written record of Zlín dates back to 1322, when it was a center of an independent feudal estate. Zlín became a town in 1397. Until the late 19th century, the town did not differ much from other settlements in the surrounding area, with the population not exceeding 3,000. Though historically associated with Moravian Wallachia, Zlín stands at the corner of three historical Moravian cultural regions; Wallachia, Moravian Slovakia and Hana [9].

In 1894, Tomáš Baťa founded a shoe factory in Zlín. The town has grown rapidly since that time. Baťa's factory supplied the Austro-Hungarian army in World War I as the region was part of the Austro-Hungarian Empire. Due to the remarkable economic growth of the company and the increasing prosperity of its workers, Baťa himself was elected Mayor of Zlín in 1923. Baťa designed the town as he saw fit until his death in 1932, at which time the population of Zlín was approximately 35,000.

Tomáš Baťa decided to sell his business to his brother Jan Antonin on May 10, 1931. Jan finished many of Tomáš' dreams and plans by more than doubling the size of the business in Czechoslovakia (in fact nearly tripling the business to nearly 50,000 people in Czechoslovakia alone). Jan Antonin was able to build dozens of city towns around the world in a less than ten year time-span.

The Bata Shoe Company reached its greatest development in 1938, the year before World War II. Jan Antonin was forced to flee from Czechoslovakia after the invasion by the Nazis. Tomas' son Thomas Jr. was Purchasing Department Manager of the English Bata Company and was unable to return until long after the war because the Baťa Company was nationalized.

The 3D visualization of Baťa Company that is described in this paper only refers to 1938. The first reason is the aforementioned greatest development period. The second reason is that there is a relatively large amount of historical material. These materials are historical photos, as well as most of the building construction plans, from which is possible to make a virtual reconstruction of this company's premises.

For 3D modeling, texturing and rendering, we used the Blender software suite [3]. It supports the entirety of the 3D pipeline— fast and effective modeling and rigging, rich animation tools, amazing simulations, photorealistic rendering, fledged compositing and motion tracking, even video editing and game creation. All textures were created in GIMP [6]. GIMP is a program under the GNU license. It is a freely distributed piece of software for such tasks as photo retouching, image composition and image authoring. It works on many operating systems and in many languages.

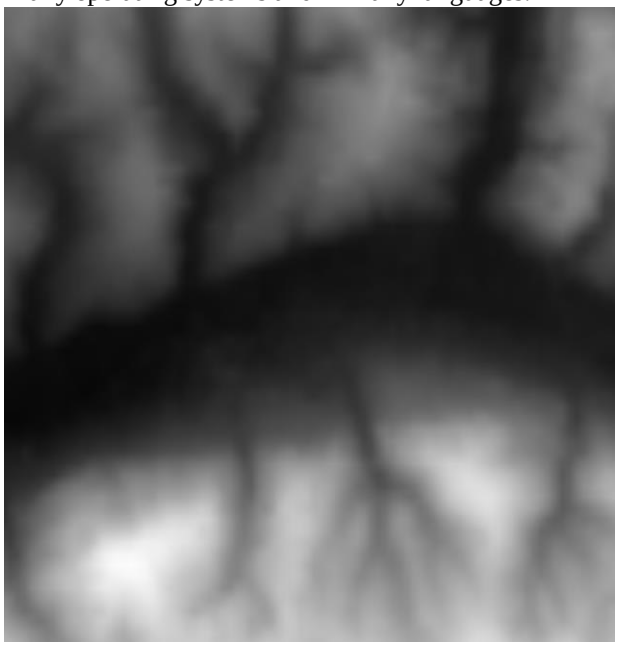


Fig. 1 the heightmap of the Zlín's Region

ISBN: 978-1-61804-236-1

II. A Landscape Model of Zlín

We used data files which contained text information about the earth elevations to create the landscape model of Zlín and its vicinity. We used the Digital Elevation Model data (i.e. DEM) which was provided by the NASA Shuttle Radar Topographic Mission (SRTM) in 2007. The data for over 80% of the globe is stored on [2] and can be freely downloaded for noncommercial use.

So we downloaded the data of Zlín's region, and then we opened it with the Microdem [7] freeware program. This software is able to convert the obtained data into a bitmap image. Microdem can clip and convert these images to grayscale (e.g. into a heightmap). We applied that to the Zlín's region. Specifically, we created a heightmap area of 25 km2 (a square with side lengths of 5 km, centered on the center of Zlín – Figure 1).

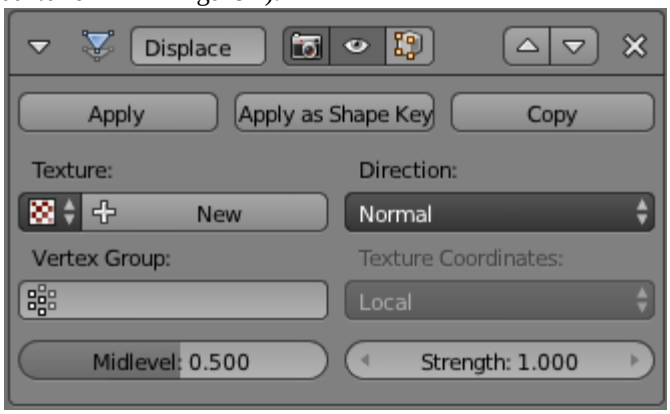


Fig. 2 Settings for the Displace modifier in the Blender environment

We saved this heightmap in PNG format (it is very important to use lossless compression). In Blender, we inserted a square (the Plane object) in the new scene and we divided it several times with the Subdivide tool to get a grid with a density of several thousand vertices. Then, we used the Displace modifier, to which we assigned a texture (for the obtained heightmap). The Displace modifier deforms an object based on the texture and setting parameters (Fig. 2). We got the model of the Zlín's landscape using this method. Although this model is not entirely accurate, but given the scale, the whole scene and the quality of the factory area model buildings, it is sufficient.

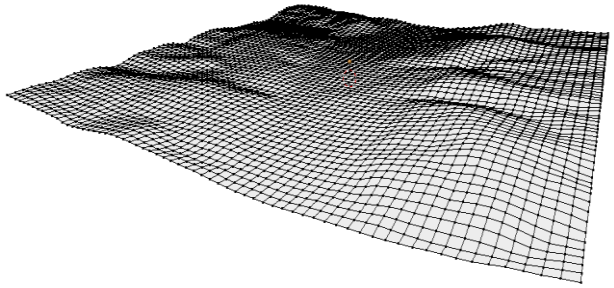


Fig. 3 the mesh model of the Zlín's lanscape

The final wireframe mesh model of the Zlín's landscape is shown in Figure 3.

III. Modeling

We began by sorting the obtained construction plans, photos and images by individual buildings before the modeling. We got these resources from [5] [8] and [10]. After that, we started the modeling. We got the ground profiles of the buildings from the period's cadastral maps; the heights and shapes were obtained from the construction plans or photos. Where these resources were missing, we improvised on the basis of other plans and photos. The level of improvisation reached around 30% in a number of buildings (the total number of factory buildings was 69).

The same problem was with the precision positions of buildings. We could not obtain the precise map of the Bat'a Company's Factory Premises in 1939. But we found the map from 1934 in the archive [8]. This map is shown in the Figure 6. It could be used for the positioning of buildings which existed in 1934. And the buildings which were constructed in 1934-1939 were positioned with the help of existing photos.

Before the modeling phase, we put the appropriate construction plans on the background screen in Blender. If these plans were missing, we used the building plans of another building with a similar shape. In these cases, based on period photos, these building plans were modified as necessary in GIMP.

We used a standard polygonal representation for the models of the buildings. Blender supports a large number of modeling tools for these so-called "mesh objects" - basic editing commands, transformation tools, modifiers, Extrude, Knife, Spin, etc. [1]

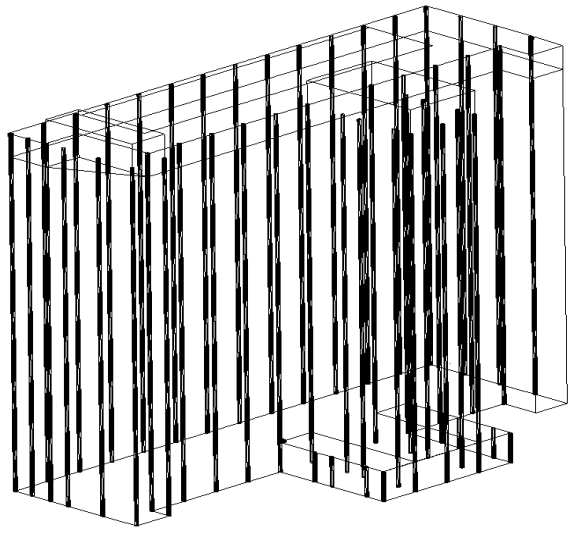


Fig. 4 the mesh model of one of factory building (wireframe shading)

The advantage was that the buildings usually have a box shape. So modeling was not difficult for this reason. To begin with, we traced the shape of the top view of the construction

plan and after that we extruded it into the third dimension. Subsequently, we modeled more specific shapes, like a rounded roof or supporting columns. An example of one building in wireframe shading is shown in Figure 4.

It is important to make "pure" models, i.e. that the models do not have unnecessary vertices, edges and surfaces and their number is viable. We have managed to reach approximately 44 000 vertices for the entire scene (68 models of buildings and accessories).

IV. Texturing and Rendering

The actual texturing was performed by UV mapping. In Blender, the command “Unwrap” transforms the 3D object surfaces into sub-surfaces into a 2D image. In this image, the borders of these faces are clearly visible; therefore, we can simply save it a bitmap format (e.g. .png in the selected resolution - we used 512x512 or 1024x1024 pixels). The next step was to open it in GIMP. It is possible to know what area belongs to which part of a 3D object according to the border, and we are also able to draw it. We used standard GIMP drawing tools and filters (Pencil, Paintbrush, Bucket Fill, Clone, etc.) or textures for their creation, which we obtained from the following reference [4] and subsequently modified them as necessary. Once the textures are drawn, they are saved in .jpg format and re-loaded in Blender, where they are properly mapped into a 3D object. We proceeded in this way for each building. An example of the uv texture of one of factory building is shown in Figure 5.

Fig. 5 the uv texture of the one of factory buildings

After the finish of the modeling phase of separated buildings, we imported each of them into the one complex 3D scene with the landscape model. In addition, we set other suitable parameters like surroundings with the sky texture; environment lighting and rendering (render resolution, anti-aliasing, shading, output raster graphic format). The last step was to find suitable camera positions and orientations in order to capture interesting render results. One of them is shown in Figure 7.

V. Conclusion

This paper describes the 3D models of the Bata factory area in 1938. This year was selected due the existence of the large amount of historical materials and, at the same time, the Bata Company was the largest development in Zlín. At this time, the whole company area included approximately 70 buildings. These buildings were modeled and textured in Blender and the final complex scene was created with rendered images and animation.

The next step is to model and texture a wider area around the factory which was closely connected to the whole operation. These items include hostels for Bata’s workers, the power plant, the sports stadium, the theater, the Moskva hotel or the shopping complex.

References

[1] R. Hess, *Blender Foundations – The essential Guide to Learning Blender 2.6*. Focal Press, 2010.

[2] A. Jarvis, H. I. Reuter, A. Nelson, and E. Guevara. (2008). Hole-filled seamless SRTM data V4, International Centre for Tropical Agriculture (CIAT) [Online]. Available: http://srtm.csi.cgiar.org

[3] Blender contributors. (2014). blender.org - Home [Online]. Available: http://www.blender.org

[4] CGTextures contributors. (2014). CG Textures – Textures for 3D, graphic design and Photoshop! [Online]. Available: http://www.cgtextures.com/

[5] FA. BAŤA A. S. (1926-1938). Fond 223 OU-ONV Zlín: plány budov, fotoarchív. Available: State District Archives Zlín – Klečůvka.

[6] Gimp contributors. (2014). GIMP - The GNU Image Manipulation Program [Online]. Available: http://www.gimp.org

[7] Blender contributors. (2014). blender.org - Home [Online]. Available: http://www.blender.org

[8] Soka Zlín. (2012). Státní okresní archív Zlín, Moravský zemský archív v Brně [Online]. Available: http://www.mza.cz/zlin/

[9] Wikipedia contributors. (2014). Zlín, Wikipedia, The Free Encyclopedia [Online]. http://en.wikipedia.org/wiki/Zlín

[10] Zlín contributors (2012). http://www.zlin.estranky.cz/ [Online]. Available: http://www.zlin.estranky.cz

Fig. 6 the original map of the Baťa Company's Factory Premises in 1934 [8]

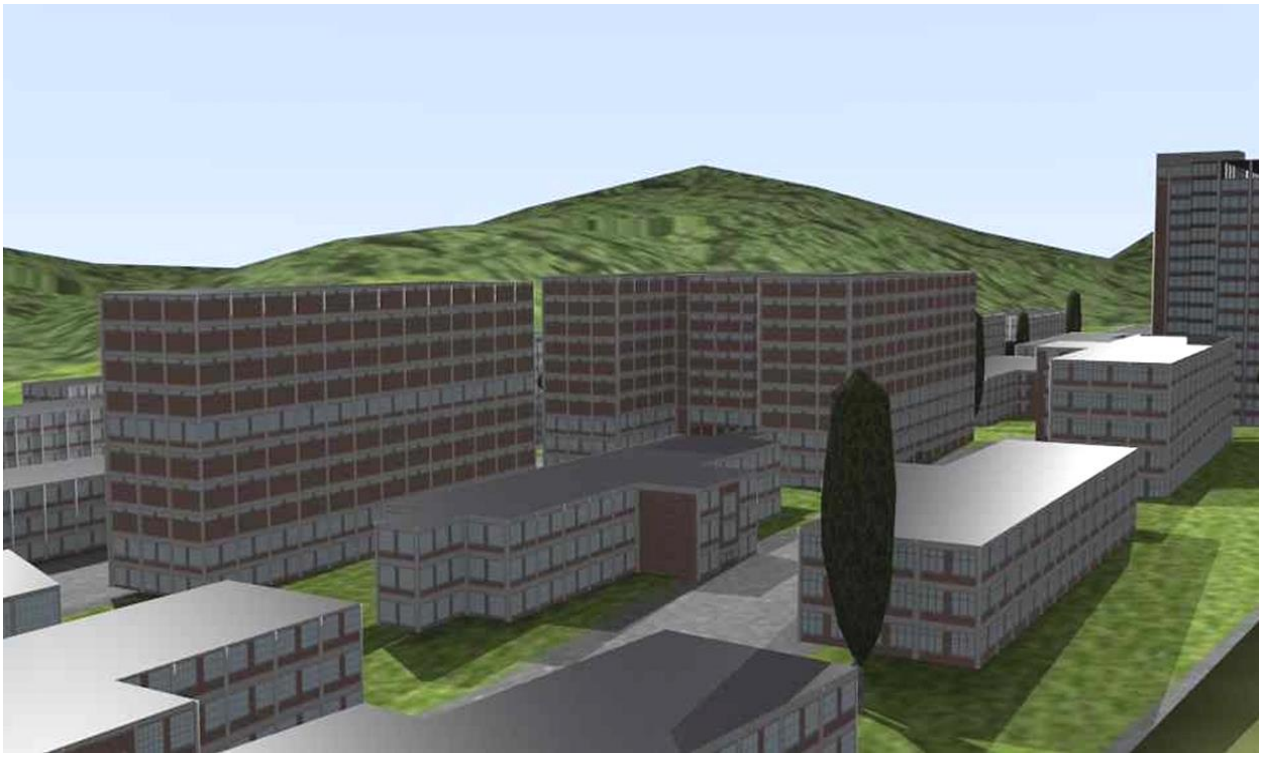


Fig. 7 the rendered image of the Baťa Company's Factory Premises model in 1939

Ordered hash map: search tree optimized by a hash table

Petar Ivanov, Valentina Dyankova, Biserka Yovcheva

***Abstract*—**An integration of a hash table and a balanced search tree is proposed. The new data structure, called *ordered hash map*, supports logarithmic-time inserts and constant-time finds and erases. An open source STL-style C++ implementations are developed which make the existing applications extremely easy to get optimized by simply changing the underlying ordered class. Our *ordered hash map* is qualitatively and quantitatively compared to the heavily used *map* and *unordered_map* STL classes.

***Keywords*—**Algorithm complexity, Binary search tree, Hash.

I. Introduction

Information technologies nowadays demand not only on storing big data but also logically organizing the information enabling fast access. Sorting, or ordering, is one of the most heavily used ways of organizing information which allows search operation to be performed fast – typically in logarithmic-time of the number of stored elements. On the other hand hashing is a technique which allows constant-time searches but does not maintain the elements ordered. Our aim is to join the ordered and hashed approaches to obtain a data structure providing constant-time search operations while keeping the elements ordered for eventual traversal. Such a structure can be used for optimizing the execution time of existing applications performing many search queries in a dynamically ordered structure.

Sorting could be done by comparison the sorted elements or by other means (radix sort, counting sort, etc.). There are approaches of ordering a hash table [1] for faster search queries but without a fast traversal over the ordered elements. This paper considers the case of sorting by comparison.

II. The Data Structure

A. Preliminary definitions

We will call a *key* an object of a specific *key type* used as an argument of a mapping. The keys of a set are called *comparable* iff a total order relation on the set is defined. The keys of a set is called *hashable* iff every key can be hashed. A *data* is any object of a specific *data type* (also called *mapped type*) which is mapped by a key.

A *self-balancing binary search tree* [2], or simply a *tree*, is a binary search tree that automatically keeps its height small (for example: AVL tree, Red-black tree, etc.). A tree can be used for mapping comparable keys to data.

A *hash map* (or *hash table*) [3] is a data structure that can map hashable keys to data.

B. Ordered hash data structure

The introduced data structure *ordered hash* consists of a tree *T* which maps keys to data and a hash map *H* which maps keys to address pointers of elements of *T*. The main reason of integrating a hash table to a search tree is to speed-up the find operation while maintaining dynamically the order of the elements in the tree.

The main operations on the introduced ordered hash:

1) ***insert(key, value)*:**
   - inserts the *(key, value)* pair into *T* as an element pointed by some pointer *p*.
   - inserts the *(key, p)* pair into *H*.
2) ***find(key)*:**
   - searches for *key* in *H*.
   - returns a pointer to *key* in *T* if *key* is found in *H*.
3) ***erase(key)*:**
   - erases *key* from *H*.
   - erases *key* from *T*.
4) ***next(tree pointer)* / *prev(tree pointer)*:**
   - returns the *next*/*prev* pointer of the given *tree pointer* provided by the tree structure.

C. Implementation

The proposed C++ class *ordered_hash_map* is fully STL-consistent [4]. It has the same interface as STL *map* so as to be easily interchanged in existing projects. The class *ordered_hash_map* is templated by *key_type* and *data_type*. The type *ordered_t* of the tree *T* is STL *map<key_type, data_type>* and the type of the hash table *H* is the STL *unordered_map<key_type, ordered_t::iterator>* (note that the map specification guarantees that elements do not change their address).

The overhead memory usage of this implementation against map is *O(N*sizeof(key))* because *H* also contains the keys. The overhead is further lowered to *O(N)* by calculating the hash function on a temporal key.

The described *ordered_hash_map* class is templated by *Key*, *Data*, *Compare*, *HashFcn*, *EqualKey*, *MapAlloc* and *HashAlloc*. This class can be naturally extended to the analogous *ordered_hash_multimap*, *ordered_hash_set* and *ordered_hash_multiset* classes.

D. Complexity analysis

All the operations on STL *map* are available on *ordered_hash_map* with the same time and memory

All the authors are from the Computer Science department of Shumen University, 115 Universitetska str., 9700 Shumen, Bulgaria.
Petar Ivanov (corresponding author) – peter.ivanov89@gmail.com.
Valentina Dyankova – valentina.dyankova@gmail.com
Biserka Yovcheva – bissy_y@yahoo.com

complexities except for the fast *find* operation which runs in *O(1)*. Note that if the number of elements *N* to be inserted is not known ahead, the complexities of insertion and erasion in hash tables are only amortized *O(1)* over multiple calls because of hash table resizing [5].

The insertion into *ordered_hash_map* consists of insertion into *T* for *O(logN)* and insertion into the *H* for *O(1)*.

The transitions to the next and to the previous element in the defined order are done by traversing *T*. The maximum edge-distance between sequential elements is *O(logN)* but the average distance is *O(1)* amortized over iterating the whole tree (as every edge is traversed exactly twice – downwards and upwards).

	map	ordered hash map	ordered hash map*	unordered map
insert	logN	logN	logN*	**1**
find	logN	**1**	**1**	**1**
erase	logN	logN	**1**	**1**
next/prev traversal	**logN**	**logN**	**logN***	n/a or N
ordered traversal	**N**	**N**	**N***	n/a or NlogN

E. Erase optimization

Instead of erasing an element from both *H* and *T*, it can be only erased from *H*. Searching for an element will be considered unsuccessful iff the element is not found in *H*. This routine optimizes the erase execution time to *O(1)*. Some additional work (not increasing complexity) is to be done when traversing the tree by the *prev*/*next* operations: every element has to be checked on whether it was not erased from *H*. This is done by using *filter iterators* from Boost library [6].

Another implication of not physically erasing elements is that the iteration time is no more dependent on the number of elements *N* logically contained in the structure but on the number of elements *N** inserted ever before. *Ordered_hash_map* with the erase optimization is referred followed by a star – *ordered_hash_map**.

F. Next/prev optimization

It is also possible to lower the complexity of the transitions to *O(1)* by maintaining two additional pointers (for the next and the previous elements) for every element in *T*. This modification increases the time constant of the implementation while not lowering the expected number of opearations in a next/prev transition. This optimization is not implemented in *ordered_hash_map* as well as in the STL *map* containter.

III. Experiments

A. Artificial data

One million uniform random string containing 100 latin latters each are mapped to random integers. All strings are sequentially inserted to a data structure, then all the strings are searched against the full data structure, then all the strings are traversed in the sorted manner, and finally all the strings are sequentially erased. The table below compares the total execution times for multiple operations execution on each of the four data structures: STL *map*, *ordered_hash_map*, *ordered_hash_map** and STL *unordered_map* (hash table). The best execution times are highlighted. As expected, *unordered_map* inserts are much faster than inserting into trees because of the constant complexity. The slight difference of the find operation executions for *ordered_hash_map* and *ordered_hash_map** is due to the construction of a filter iterator. The slower *ordered_hash_map** traversal is due to the additional filter iterators.

	map	ordered hash map	ordered hash map*	unordered map
insert	2.68	3.60	3.63	**0.83**
find	2.67	**0.51**	**0.63**	**0.49**
erase	3.32	4.05	**0.94**	**0.91**
next/prev traversal	**0.21**	**0.22**	0.94	n/a

IV. Applications

Ordered hashes can be used instead of STL maps/sets for all kinds of hashable elements. Giving a hash function is not needed if STL containers are used as elements.

Sensitive hashing [7] and some other kinds of locally sensitive hashes are of a particular interest because of their inherited presumption for sorted hash values.

V. Conclusion

We hope that the presented data structure will be useful in practice because of its standard interface and its better complexity and find execution speed compared to the most commonly used alternatives. The future work includes testing the data structure on more real test cases.

All described C++ classes and detailed experimental results are available with open source licensing at https://github.com/petar-ivanov/ordered-hash.

VI. Funding

This paper is supported by the Project BG051PO001-3.3.06-0003 “Building and steady development of PhD students, post-PhD and young scientists in the areas of the natural, technical and mathematical sciences”.

References

[1] O. Amble and D. E. Knuth, Ordered Hash Tables, The Computer Jurnal (Oxford, 1974), Volume 17(2).

[2] D.Knuth, The Art of Computer Programming, Volume 3: Sorting and Searching, Second Edition, Addison-Wesley, 1998, Section 6.2.3: Balanced Trees, pp.458—481.
[3] T. Cormen, C. Leiserson, R. Rivest, C. Stein, Introduction to Algorithms (2nd ed.), MIT Press and McGraw-Hill, 2001, pp.221—252.
[4] SGI Standard Template Library (STL) Programmer's Guide, Available: https://www.sgi.com/tech/stl/
[5] C. Leiserson, Amortized Algorithms, Table Doubling, Potential Method, Lecture 13, course MIT, Introduction to Algorithms – Fall 2005.
[6] D. Abrahams, J.Siek, T. Witt, Boost Filter iterators, Available: http://www.boost.org/doc/libs/1_55_0/libs/iterator/doc/filter_iterator.html
[7] C. Sadowski, Greg Levin, SimHash: Hash-based Similarity Detection

Comparative Analysis on the Competitiveness of Conventional and Compressive Sensing-based Query Processing

Salema Fayed[a], Sherin Youssef[a], Amr El-Helw[b], Akbar Sheikh Akbari[c], Mohammad Patwary[c] and Mansour Moniri[c]

[a] Computer Engineering Department, [b] Electronics and Communication Department, College of Engineering and Technology, Arab Academy for Science and Technology, Alexandria, Egypt

[c] Faculty of Computing, Engineering and Technology Staffordshire University, Stoke on Trent, United Kingdom

***Abstract*—**Optimization of the lifetime of the battery within wireless sensor networks (WSNs) is challenging due to communication infrastructure. Subsequently, minimizing the amount of power required for data collection and processing to serve the intended purposes has become an open research problem. Conventional and compressive sensing-based (CS) query processing being the candidates to perform these tasks, require a comparative analysis in the current WSN application context. In this paper. Simulations have been carried out to compare the performance of conventional and compressive sensing-based (CS) query processing with respect to energy efficiency, sensing reliability and normalized estimation error within WSN. A significant reduction in the computational complexity reaching 70% is noticed using CS compared to conventional query processing algorithms. Moreover, it is observed that up to 90% sensing reliability can be achieved with CS compared to existing query processing. Hence, the reduction in computational complexity has not compromised the sensing reliability with an observed reduction in the normalized estimation error.

***Keywords*—**WSN, Compressive sensing, Query processing

I. Introduction

Wireless sensor networks are one of the first real-world examples of pervasive computing [1]. Sensor networks [2],[3] consists of small, low-cost, limited battery-operated sensors, which collect and disseminate environmental data from remote locations without human interaction for weeks or months at a time. One of the major challenges in designing WSN is to maximize network lifetime which requires the network to be energy efficient by reducing the communication cost and processing complexity to enable operation for extended duration. Hence, power conservation and power management take on additional importance. Unlike other environments, query-processing systems designed for sensor networks must integrate energy awareness into the system to extend the lifetime of the sensor nodes and network. In consequence, the network should be self-configured in order to reduce the energy consumption without significantly diminishing the coverage and connectivity of the network [4]. Different Query processing techniques have been investigated. There has been some related work in the area of sensor network's self-configuring as the Span algorithm proposed in [5]. It has the potential for significant reduction of message loss and increase in energy efficiency. Although this has resulted in energy saving but investigations have shown that higher density can be extremely expensive in terms of broadcast messages keeping nodes awake for a longer time. Some existing query processors such as acquisitional query processing (ACQP) [2] have met these requirements by pushing operations such as selection and aggregation within the sensor network in order to reduce the communication resource and processing cost. However, their approach has some deficiencies with respect to query optimization and routing.

In [6], the authors have proposed several decentralized query- processing approaches to utilize sensor node's innate spatial and semantic characteristics. The algorithms faced some problems in terms of high consumption of energy in query optimization; query plans are still not guaranteed to be efficient for all nodes. Rather than using all the sensors all the time to monitor events and queries, set-K-Cover solutions [7] provide a simple way for sensors to share in the monitoring of an event. In [4], the authors have proposed an algorithm for efficient query processing, where the network is self-organized in response to a given query. Still the query processing requires higher message overhead and higher computational complexity. While in [8] a near-optimal query processing algorithm has been introduced, which results in a significant reduction in the energy consumption and message overhead, yet smart sensors are needed for the processing which involve extensive computations. For these reasons, there is much research in the design of energy- efficient protocols and algorithms for sensor networks.

ISBN: 978-1-61804-236-1

A new research area known as compressive sensing (CS) first proposed by Candes in [9] for data acquisition and processing. It senses the networks environment without the need of smart processing and intelligent decision-making at each sensor node. Furthermore, CS does not require the extensive computation as in WSN query processing algorithms [9], [10]. In [11] a framework is proposed to adaptively collect information from a WSN using adaptive compressive sensing taking into account both energy consumption and the amount of information in the sensing data. In [12] an energy-efficient compressed sensing for data gathering in WSN using spatially-localized sparse projections has been proposed. The designed method has resulted in power saving over other existing techniques. Hence, CS could be a strong candidate to achieve energy efficient data gathering in WSNs requiring simple and low power computations. In this paper, a comparative analysis on the competitiveness of CS and traditional query processing in WSNs is carried out. The performance of two different optimized query processing techniques [4],[8] and CS has been compared with respect to energy efficiency, sensing reliability and normalized error.

The rest of the paper is organized as follows, candidate query processing techniques investigated for comparison are briefly summarized in Section II. Compressive sensing in WSN is presented in Section III. Section IV presents the system model. Competitiveness analysis and performance indicators have been identified in Section V. Simulation results have been provided in Section VI and finally the conclusion in Section VII.

II. QUERY PROCESSING IN WSN

Query processing for WSN has been investigated in many literatures, such as [2], [4]–[8]. There is always a tradeoff between energy efficiency, communication cost, computation complexity and accuracy hence for a fair comparison, high performance query processing techniques are selected for the competitiveness analysis; the energy-efficient distributed self-organization algorithm for near-Optimal sensor cover (EED-SOSC) [8] and the greedy self-organizing decentralized approximation algorithm (GDAA) [4]. These candidate algorithms are summarized below:

A. The Energy-Efficient Distributed Self-Organization Algorithm for near-Optimal Sensor Cover (EEDSOSC):

EEDSOSC algorithm is applied for query processing presented in [8], it has been proven that it produces a near-Optimal Sensor Cover with minimum energy consumption, minimum cover- size and less message overhead. The objective is to select a connected sensor to provide the predefined coverage with the minimum total energy consumption.

For a given query about a region *QR*, it is desired to find a set of sensors covering *QR*, this set of sensors is denoted as the cover *C*, for the selection of candidate sensors, the control is distributed over the set *B* of boundary sensors around *C* as shown in Fig.1. Initially, a pool of candidate initial sensors is constructed with sensors with longer lifetime (i.e. lower consumed energy) is to be included in the initial pool. The algorithm can be summarized with the following steps:

- Step 1: Choosing the boundary sensors that fall at a threshold distance from the center of the cover.
- Step 2: Finding possible Candidate sensors C_s via a view candidate sensor (VCS) message.
- Step 3: Collecting Candidate paths C_p connecting C_s to sensors in B
- Step 4: Choosing the most beneficial candidate path/sensor $\varpi = A_{C_S}/W$, where ϖ is the benefit calculation of each C_s, A_{C_S} is the sensing area of the corresponding sensor C_s and W is the energy weight associated with each sensor.
- Step 5: Check for query coverage.

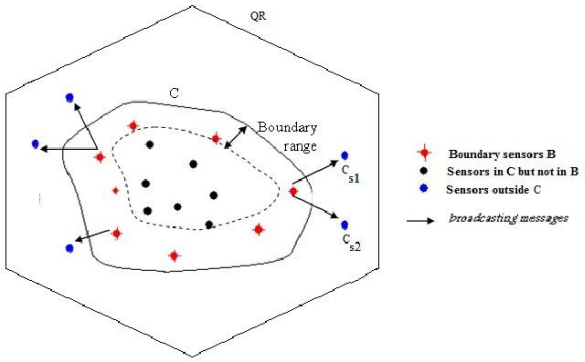


Fig. 1 Model for the EEDSOSC algorithm

B. Energy-efficient greedy decentralized approximation algorithm for query processing (GDAA):

This algorithm has been presented in [4]. In brief, to build a connected sensor cover within the sensor network for a given query, at each stage the greedy algorithm works by selecting a path (communication path) of sensors that connects an already selected sensor to a sensor of partial coverage. The selected path is then added to the already selected sensors at that stage. The algorithm terminates when the selected set of sensors completely cover the given query region. But at each stage, newly added sensors are added via the most recently added sensor from the previous stage, for that reason the algorithm takes longer time to terminate as the solution is sometimes deviated away from the desired region.

The algorithm EEDSOSC is based on a fully decentralized selection criterion while GDAA is based on a greedy selection as the new selection depends only on the recently selected sensor C_s^* in the cover C whereas EEDSOSC depends on all sensors falling on the boundary of C. The algorithm can be summarized with the following steps:

- Step 1: Finding possible Candidate sensors C_s via a view candidate sensor (VCS) message.
- Step 2: Collecting Candidate paths C_p connecting C_s to sensors in C

- Step 3: Choosing the most beneficial candidate path/sensor $\varpi = A_{C_s} / N_{C_p}$, where ϖ is the benefit calculation of each C_S, A_{C_S} is the sensing area of the corresponding sensor C_S and N_{C_p} is the number of sensors in the selected path.
- Step 4: Check for query coverage

III. COMPRESSIVE SENSING IN WSN

A. Overview

Compressive sensing (CS) is a new paradigm for data acquisition and processing, it was originally developed for the efficient storage and compression of digital images [9], [13]. Theory of CS as presented in [9] claims that signal can be reconstructed with far fewer samples than that required by the Nyquist theory. However, sparse nature of signals, where most of the signal's energy is concentrated in few non-zero coefficients is to be provided. Furthermore, it is not necessary for the signal itself to be sparse but compressible or sparse in some known transform domain Ψ. According to the nature of signals as an example, smooth signals are sparse in the Fourier basis, and piecewise smooth signals are sparse in a wavelet basis [14].

Suppose a signal s of size N is represented with fewer samples instead of all the elements by random projection through incoherent measurements m of size M ($M << N$), where M is constrained by (1)

$$M \geq K \log N \tag{1}$$

and K is the number of non-zero coefficients, the measurements m is calculated as in (2)

$$m = \Phi s \tag{2}$$

where Φ is a random matrix of size (M x N). Φ must obey uniform uncertainty principle (UUP) and thats if $M \geq K \log N$ [14], [15], this guarantee reconstruction of the signal/image [10]. Φ is unstructured and universally incoherent to Ψ and with every measurement fractional information about the sparse coefficients can be obtained. Hence, it does not have to match any structure of the signal but to look more like random noise than any feature of the signal matrix [9].

IV. SYSTEM MODEL

For a unified comparison among the candidate processing techniques mentioned in previous section, a generalized sensor network communication model is designed as shown in Fig.2. A wireless sensor network for sensing and monitored is modeled for a geographical region of dimension (X x Y) square unit. Such geographical region is assumed to be divided into G number of grids. Hence the dimension of each grid is (g x h) square units, where $g = X/G$ and $h = Y/G$. N sensors are randomly distributed (uniform random) within the network's coverage region (X x Y). It is also assumed that, the dimension of the grid is chosen in such a way that one sensing node is enough to cover a single grid for monitoring purposes. Assuming omnidirectional antenna for each of the wireless sensing nodes, the typical coverage radius is defined by (3):

$$r_s = \frac{1}{2}\left(\sqrt{g^2 + h^2}\right) \tag{3}$$

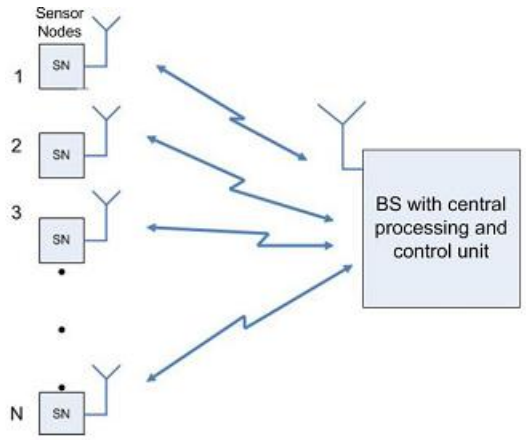


Fig. 2 A unified WSN communication model

The monitoring activity factor within the sensing environment is defined by η, expressed as:

$$\eta = \frac{N}{G} \tag{4}$$

All sensor nodes are controlled from a base station (BS) with central control and processing unit. Suppose each sensor node has a unique identifier and the coordinates of all the sensors are known to the central system. Besides that, each sensor node is able to communicate with its neighboring nodes if they are at a distance apart within the specified transmission range t_r, where $r_s \leq t_r \leq 2r_s$. Assume that initially all sensor nodes have uniform battery life. However, cumulative power consumption of each sensor differs from the other during its lifetime. The battery life of sensor nodes depends on several factors and accordingly [16], power consumed by each sensor node is P_i can be expressed as (5):

$$P_i = \frac{k w_i Q}{d_i}\left(1 + \frac{\Delta\mu_n}{\delta}\right) T_{avg} \tag{5}$$

Where, k is the proportionality constant, δ is the sensitivity threshold of detecting the occurrence of an incident, ($\Delta\mu$) is the step size of the sensing gradient, n is the number of steps, d_i is the radial distance of the sensor from the center of the incident, w_i is the frequency of occurrences of incidents, Q is the computational operations performed at each sensor node, and T_{avg} is the average period of the incident. The time T_{avg} is modeled in a time-

stepped fashion, wherein during each step, each node receives messages, perform appropriate computations in response to these message and sends out messages as a result. Hence, the sensor battery life in terms of the number of responses to the incidents is defined as in (6):

$$\beta = \frac{E_B}{P_{avg} T_{avg}} \tag{6}$$

Where, E_B is the energy rating of the battery and P_{avg} is the average power consumed by sensor node, $P_{avg} = E(P_i)$. Sensor nodes are queried by the BS to monitor relevant data about the surrounding environment. In response the sensor node is required to report the monitored incidents back to the BS for decision making, hence reliable sensing is required from the node. Sensing reliability depends on the delay, computational complexity at sensor nodes. Let's assume that each sensor has a probability of α to detect an incident if the incident is within its sensing region. i.e. the sensing probability is expressed as in (7).

$$p_{vi}(x,y) = \begin{cases} \alpha & if\ d_i \le t_r \\ 0 & otherwise \end{cases} \tag{7}$$

Where $0 \le \alpha \le 1$; v_i is the i^{th} sensor at coordinates (x, y). Given a query over a sensor network, it is required to select a set of M sensors that is sufficient to represent the physical environment energy-efficiently and reliably. If an incident is covered by sensing nodes vector [v_1, v_2, . . . , v_M], its sensing probability is defined by (8):

$$p_{v_1,v_2,\ldots,v_M}(x,y) = 1 - \prod_{i=1}^{M} (1 - p_{vi}(x,y)) \tag{8}$$

and the joint sensing probability (JSP) for these M sensors is given by (9):

$$E(p_{(v_1,v_2,\ldots,v_M)}) = \int\int p_{v_1,v_2,\ldots,v_M}(x,y)\beta dxdy \tag{9}$$

To maximize the lifetime, the total power consumed by each sensor node during the response of any incident is required to be minimized. Moreover, any reduction in the computational complexity and in the number of messages exchanged is expected to result in reducing the energy consumed by the sensors.

Another parameter to measure the performance of reconstructing the signal sensed by nodes is the normalized estimation error which can be derived according to [15] as (10)

$$\varepsilon = \frac{\Omega}{q\rho} \tag{10}$$

where $\Omega = 2BW$ represents the minimum Nyquist sampling rate and BW is the sensed signal's bandwidth; q is the probability of correct sample transmission; ρ is the density of sensor distribution.

V. COMPETITIVENESS ANALYSIS

In the following subsections, CS in WSN is compared with previously mentioned query processing algorithms EEDSOSC and GDAA. The performance indicators is in terms of power consumption by sensor nodes, sensing reliability, and normalized estimation error in response to a given incident. Afterwards, the results of the simulations are presented..

A. Power consumption

From (5), it is obvious that all parameters affecting the power are the same in all techniques except for Q. The computations at sensor nodes differ according to the complexity of each technique. Subsequent sections shows the calculations of Q for the three techniques.

1) EEDSOSC: For a network of N nodes, the number of computational operations Q_{ed} in terms of additions and multiplications is calculated as in Table.I. Where, $D = C_2^N$ distance calculations of the Euclidean distance d_{ij} between node i and node j; $\vartheta_{ed}{}^a$ and $\vartheta_{ed}{}^m$ are constants representing additions and multiplications, respectively required throughout the rest of the algorithm. $\vartheta_{ed}{}^a \approx \boldsymbol{I}[17S + 4V + VS]$ and $\vartheta_{ed}{}^m \approx 3\boldsymbol{I}[S + V]$, where $\boldsymbol{I}$ is the number of iterations needed for the algorithm to converge, S is the number of candidate sensors at each iteration and V is the number of sensors in the query coverage C such that $S,V>1$.

2) GDAA: As in sec.V-A1, the Q_{gd} is calculated in similar manner, where $\vartheta_{gd}{}^a$ and $\vartheta_{gd}{}^m$ are constants representing additions and multiplications performed till the convergence of the algorithm, respectively. GDAA has gone through the same computations as in the decentralized approximation algorithm [8] except for the stage of finding the boundary sensors, and the threshold parameters. Each stage is dependent only on the recently added sensor, hence no computations are required concerning all V sensors selected within the coverage. Hence, $\vartheta_{gd}{}^a = 17IS$ and $\vartheta_{gd}{}^m = 3IS$

3) CS: Whereas in the compressive sensing technique for the same network of N nodes, the only operation required is multiplying sensor's matrix s by their coefficients ϕ of dimension (N x M). Q_{cs} is defined in Table.I.

Table 1 Number of computations Q in terms of additions and multiplications for the three techniques EEDSOSC, GDAA, CS

Q	Additiions	Multiplications
Qed	3Dϑeda	3Dϑedm
Qgd	3Dϑgda	3Dϑgdm
Qcs	(N-1)2 M	N2 M

B. Sensing reliability

Sensing reliability is measured by calculating the probability of false sensing of incidents $q_{v2,\ v2,\ldots\ vM}(x,y)$ From (8), the probability of false sensing is given by (11) where α is generated by gaussian random numbers depending on the distance between the incident and the sensor.

$$q_{v_1,v_2,\ldots,v_M}(x,y) = \prod_{i=1}^{M}(1 - p_{vi}(x,y)) \quad (11)$$

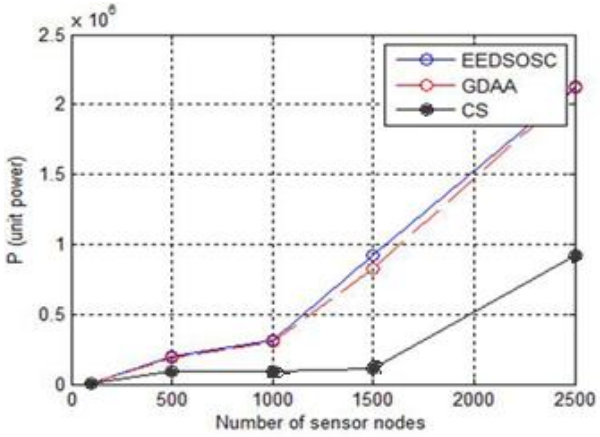


Fig. 3 Power consumption for EEDSCOSC, GDAA and CS versus different network sizes

VI. Simulations and Results

In this section, a simulator is constructed to build the comparison. The algorithms ran on uniform random generated sensor nodes, wherein a certain number of sensor nodes N are distributed in an area of $X = Y = 100$ (unit distance). N is varied from 100 to 2500 uniform random placed sensors. The size of rectangular region, number of nodes, sensing radius, and transmission range are input parameters to the simulator. For the query processing techniques, the simulator only models the transmission of messages involved in the nodes selection process.

It is illustrated in (5) that $P_i \propto Q$, to compare between Q for all techniques. According to Table. I, Fig.3 demonstrates the power consumption for these three techniques. It is obvious that there is an overlap between the two curves for the EEDSOSC and the GDAA as the difference can be negligible. It is clear from the graph, as the number of sensors increases the reduction in power consumption increases for CS compared to query processing techniques as a result of the significant reduction in Q_{cs} .

Fig.4 illustrates different probabilities of false sensing for the three comparative techniques. It is shown that for any number of sensors the probability of false sensing in CS is less than other query processing techniques, hence the sensing probability increases according to (8). Both EEDSOSC and GDAA have nearly the same probability of false sensing. While CS does not select sensor nodes compared to EEDSOSC and GDAA, it senses the network's environment through fewer measurements. Therefore, the number of sensors selected by CS is higher than the candidate query processing techniques resulting in higher sensing probability.

For the query processing techniques either EEDSOSC, or GDAA, the estimation error in reconstructing the sensed phenomena from (10) is the same, as all parameters are common to these two techniques. However, in the CS technique ε is reduced as the main characteristic of CS is that the sampling rate is much less than Nyquist rate, and samples are

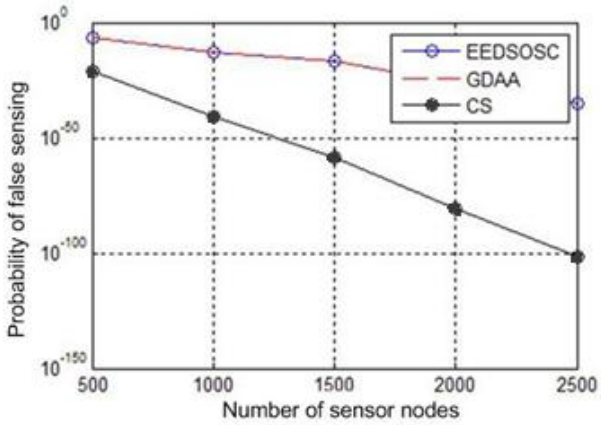


Fig. 4 Probability of false sensing for EEDSOSC, GDAA and CS versus different network sizes

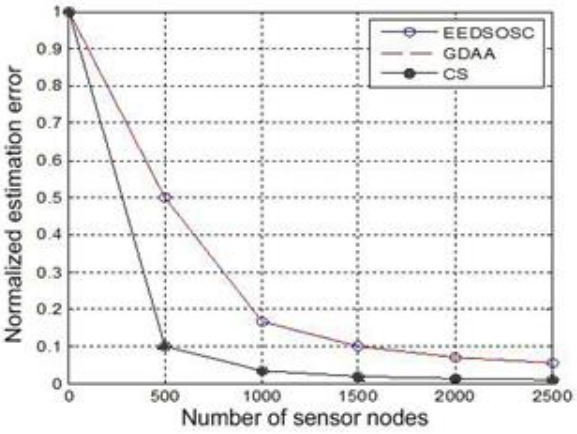


Fig. 5 Normalized estimation error for EEDSCOSC, GDAA and CS versus different network sizes

reconstructed with fewer number of measurements than double its bandwidth. Hence, in the CS $\Omega_{cs} << \Omega$ yielding smaller ε. Subsequently, as N increases, ρ increases and the normalized estimated error decreases. Fig.5 illustrates the comparison of estimation error ε for all the three techniques versus different network's sizes and densities. It is observed from the plot that the estimation error decreases as N increases. EEDSOSC and GDAA have the same normalized estimated error as both sample at Nyquist frequency. On the other hand, CS uses fewer samples compared to the Nyquist sampling that has yield to lower ε. For different network sizes, the figure depicts a reduction in estimation error using CS compared to the other techniques. The results

prove that the improvement in terms of energy efficiency has not affected the performance or reliability of the CS technique.

VII. CONCLUSION

A comparative analysis of conventional and compressive sensing-based query processing is carried out. Two conventional query processing techniques; A greedy decentralized approximation algorithm GDAA and an Energy Efficient Distributed Self-organization algorithm EEDSOSC, have been studied and compared to sensing the network via CS. The conventional query processing techniques efficiently select a set of sensors according to different criteria that are sufficient to respond to particular query. Different Network sizes have been studied to illustrate their effect on the performance indicators. Simulations have shown that CS has proven a notable reduction in the computational complexity reaching 70% compared to the EEDSOSC and GDAA.
Furthermore, CS has provided great improvement in sensing reliability by reducing the probability of false sensing by 90% compared to its two other counterparts. Besides that, information collection with compressive sensing has shown a notable reduction in the estimated error in response to an incident without compromising the sensing reliability.

REFERENCES

[1] N. Bulusu and S. Jha, Wireless sensor networks, Artech House, INC publication, London, 2005.

[2] S. R. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong, "The design of an acquisitional query processor for sensor networks," in ACM SIGMOD International Conference on Management of Data, San Diego, California, June 2003, pp. 491–502.

[3] J. Lian, L. Chen, K. Naik, M. zsu, and G. Agnew, "Bbs: An energy efficient localized routing scheme for query processing in wireless sensor networks," International Journal of Distributed Sensor Networks, ACM, vol. 2, no. 1, January 2006, pp. 23–54.

[4] H. Gupta, Z. Zhou, S. R. Das, and Q. Gu, "Connected sensor cover: Self-organization of sensor networks for efficient query execution," IEEE/ACM Transactions on Networking, vol. 14, no. 1, 2006, pp. 55–67.

[5] B. Chen, K. Jamieson, H. Balakrishnan, and R. Morris, "Span: an energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks," in Proceedings of International Conference on Mobile Computing and Networking (MobiCom), 2001, pp. 85–96.

[6] R. Rosemark and W. Lee, "Decentralizing query process- ing in sensor networks," in The Second Annual Interna- tional Conference on Mobile and Ubiquitous Systems: Networking and Services, 2005, pp. 270–280.

[7] A. G. Z. Abrams and S. Plotkin, "Set k-cover algorithms for energy efficient monitoring in wireless sensor net- works," in Proceedings Int. Workshop on Information Processing in Sensor Networks IPSN, 2004, pp. 424–432.

[8] M. A. Hamza, S. M. Youssef, and S. F. Fayed, "A distributed energy efficient query processing in self- organized wireless sensor networks," in Proceedings of the World Congress on Engineering, London, U.K., vol. 2, July 2007.

[9] E. J. Candes, "Compressive sampling," in Proceedings of the International Congress of Mathematicians, 2006.

[10] J. Romberg, "Imaging via compressive sampling," IEEE Signal Processing Magazine, March 2008, pp. 14–20.

[11] C. T. Chou, R. Rana, and W. Hu, "Energy efficient information collection in wireless sensor networks using adaptive compressive sensing," in IEEE 34th Conference on Local Computer Networks LCN, Zrich, Switzerland, October 2009, pp. 443–450.

[12] S. Lee, S. Pattem, M. Sathiamoorthy, B. Krishnamachari, and A. Ortega, "Spatially-localized compressed sensing and routing in multi-hop sensor networks," in Proceedings of the 3rd International Conference on GeoSensor Networks, July 2009.

[13] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," IEEE signal processing magazine, pp. 21–30, March 2008.

[14] R. G. Baraniuk, "Compressive sensing," IEEE Signal Processing Magazine, pp. 118–124, July 2007.

[15] A. Conti and D. Dardari, "The effects of nodes spatial distribution on the performance of wireless sensor net- works," in IEEE Vehicular Technology Conference VTC, vol. 5, May 2004, pp. 2724 – 2728.

[16] M.Tahir, N.Javaid, M.A.Khan, S.Rehman, A.Javaid, and Z.A.Khan, "Energy efficient transmission in wireless sensor network," Research Journal of Applied Sciences, Engineering and Technology, pp. 723–727, 2013.

A 3D Visualization of the Tomas Bata Regional Hospital Grounds

P. Pokorný and P. Macht

***Abstract*—**This paper describes the design and implementation of 3D visualization of the Tomas Bata Regional Hospital grounds in Zlín. In relation to historical materials which are available in the state archives and libraries, the first main aim of this work was to collect images of individual buildings from 1927 (this was the year that the Tomas Bata Regional Hospital was founded) to the present day, including its surrounding area. All the collected information was chronologically sorted and on this basis, we created a 3D visualization of the urban development of the hospital grounds. In addition to the 3D models we created (the standard polygonal representation), we designed and mapped suitable textures to get the actual appearance in the time period of individual buildings in the grounds. Using this process, we created full 3D models in these years: 1927, 1930, 1935, 1940, 1950, 1960, 1980, 1990, 2000, 2005 and 2013. The visualization is performed by rendered animations in these years.

***Keywords*—**Visualization, 3D graphics, Modeling, Texturing, Animation.

I. Introduction

THE history of the Tomas Bata Regional Hospital in Zlín dates back to May 1926, when the project was proposed. The hospital was situated approximately 3 km east of the city center on an almost square plot of land between the Dřevnice River and the forest. This forest was felled some years later and houses were built in its place, so the hospital was directly connected to the entire city.

The first hospital buildings were built in 1927 – i.e. the entrance building and two pavilions. By 1938, 16 hospital pavilions had been built, including the original buildings. Each pavilion was designed with the standardized appearance and typical architecture for most buildings in Zlín at this time – the combination of red bricks and a light gray concrete. The structure of the pavilions was also unified. Slight differences were only given by the specific needs of individual departments.

New development plans were created for the further expansion of the hospital grounds in 1946. Based on these plans, over the following years, new five pavilions were constructed. For the ensuing almost 30 years, the hospital grounds were without major changes. Only some reconstructions, adaptations and changes of pavilions were carried out over the years.

The next significant development of the hospital grounds was implemented after 1973, when the Surgery Building, Pathology-anatomical Department, District Health Station, Internal Departments and Utility Energy Block were developed in step-by-step phases. The new Ophthalmology pavilion was built in 1984 and the Hospice was built in 1989. The Tomas Bata Regional Hospital grounds encompasses nearly 60 buildings numerically designated for orientation purposes at present. These buildings have different medical and/or technical functions. Our main objective was to create complex 3D visualizations of the construction phases of the Tomas Bata Regional Hospital grounds in 1927, 1930, 1935, 1940, 1950, 1960, 1980, 1990, 2000, 2005 and 2013. The choice of these years was based on the available documentation, and these were the years that brought the greatest changes in the construction phases of the entire hospital complex.

II. Acquiring Resources

The overall progress of this work was initiated by the collation of available historic materials and information about the Tomas Bata Regional Hospital. The main resources were the State District Archive in Zlín – Klečůvka [5], the Moravian Land Archive in Brno and the Tomas Bata Regional Hospital Archive. We mainly focused our attention on building plans, cadastral maps and historical photos.

In the archives in Zlín – Klečůvka and Brno, we mainly found historical photos of this hospital from the foundation to the end of the nineteen-forties. Most resources and information were obtained in the Tomas Bata Regional Hospital Archive, where we found most of the construction plans of the individual buildings. We also held discussions with the hospital staff about the whole hospital grounds and its development over time.

The two books published for the 75th and 80th anniversaries of the Tomas Bata Regional Hospital [1], [2], were the next two important resources. These books contain information about the construction of buildings in the grounds as well as period photographs that were not stored in the archives. A website called Old Zlín [6] was the last important source in the creation of this work. The information on this webpage helped to verify some information which was obtained from the other references.

Based on the materials we obtained, we created a table that contains all of the acquired data on the construction sites,

ISBN: 978-1-61804-236-1

renovations and demolitions of individual buildings. This table was divided into several parts. In the first worksheet, we created a complete list of buildings, from their construction to the present or their possible demolition. For greater clarity, we created additional document pages that contain all construction events in the area in five-year intervals.

From this table, we created a bar graph that shows the number of newly constructed buildings over five-year intervals. This graph is shown in Figure 1.

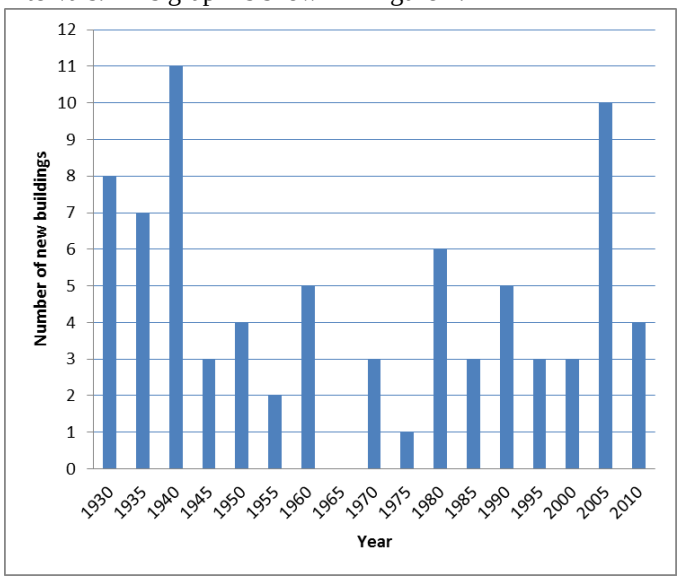


Fig. 1 number of new buildings in five-year intervals

III. Modeling

For 3D modeling, texturing and rendering, we used the Blender software suite [7]. Blender is a free and open source 3D animation suite. It supports the entirety of the 3D pipeline— fast and effective modeling and rigging, rich animation tools, amazing simulations, photorealistic rendering, fledged compositing and motion tracking, even video editing and game creation.

A. Preliminary work

During the analysis process of the materials, we found a large amount of construction documents. We placed some of them as background images into Blender and based on them we created the models of the hospital buildings. Above all, we used front, side and top view construction plans especially.

However, many of them had different scales. So we had to adjust them to the same scale. To unify the scale of all the plans we used Blender Scale tool, which can change the background images´ size in both dimensions.

The next step before starting the modeling of buildings is to select the level of the models detail.

Given the scale of the whole scene and the purpose of our visualization (the presentation of the entire hospital complex), we determined the level of detail of each building model down to the level of niches of windows. All other less "inequalities" in the objects were only drawn in the textures. The great advantages of this solution are the considerable reduction system memory consumption and thus great accelerations of the rendering process.

B. Modeling Buildings

The modeling of the buildings was always performed according to the same scenario in the Blender program. In the scene, we first put the Plane object and then we modified its profile (i.e. the size and shape) according to the floor construction plan of the building. We mainly used two tools to shape this object. The Extrude tool, which allows us to alter the selected face in a specifically chosen direction and the Subdivide tool, which breaks down the selected face into even greater number of smaller parts [3].

When the floor shape was finished, we extruded this profile to a height corresponding to the construction documents. Extrusion was performed for each floor until we reached the total height of the building.

The next step was the necessity to model the building roof. The buildings in the hospital complex have three different types of roofs – a pitched roof, a flat roof with an overhang and a flat roof with a raised edge (Figure 2). The pitched roof was created by dividing the top face into two parts, and the newly created edge was subsequently moved to the required height. The flat roof with an overhang was created by one more extrusion and the newly formed faces were pulled in the normal direction. The flat roof with a raised edge was created by the Inset Faces tool, which creates new smaller faces at the edges of the selected area. After that, these new sub-faces were subsequently extruded in the vertical direction.

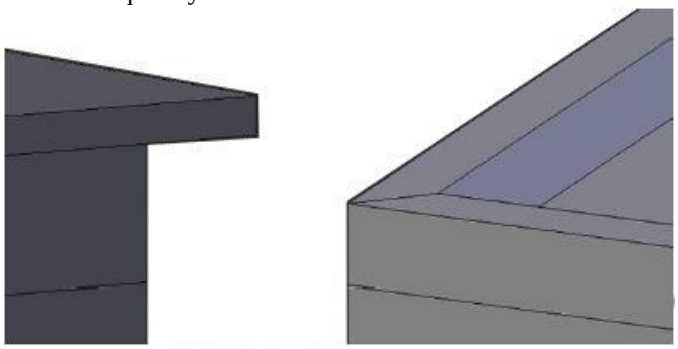


Fig. 2 a flat roof with an overhang (Left) and a flat roof with a raised edge (Right)

To make the windows and doors embedded in the buildings, we used the Subdivide tool again on the relevant faces. After this, we deleted these new sub-faces to model holes. Additionally, we extruded the border edges of these holes to the depth of embedment of windows and doors.

Accessories (e.g. windows, doors, chimneys, railings, etc.) were modeled as separate objects by using the tools described above. All sub-models of each building were subsequently linked to the main building model to unify manipulations

with the whole building after completing the model.

An example of one modeled hospital building is shown in Figure 3. In this picture we can see the 13th pavilion model from 1929 in the Blender environment.

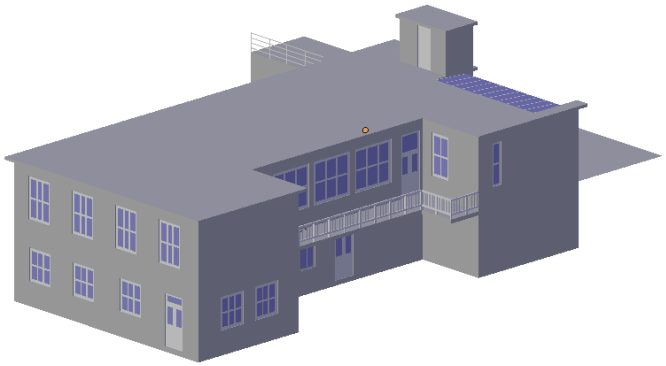


Fig. 3 the 13th pavilion model from 1929 in the Blender environment

C. *Modeling Terrain and Accessories*

Before the modeling terrain phase, we put the appropriate ground construction plan of the whole hospital area on the background screen in Blender. After that, we added the Plane object in the 3D scene and the edges of this object were shaped by the boundaries of the hospital area. We used the Knife tool to multiply the cuts made this object in the next step in order to get a lot of new vertices. In the next step, these vertices were placed in suitable positions (i.e. on most roads and paths between hospital pavilions) in order to create the most accurate possible 3D terrain model (Fig. 4).

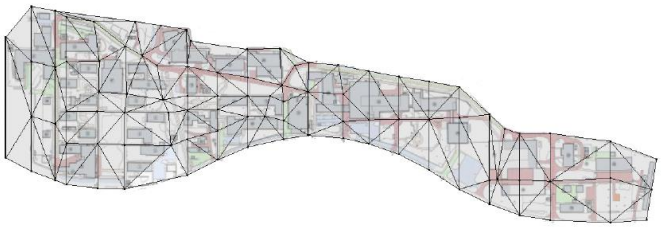


Fig. 4 the ground hospital plan with the inserted ordered vertices of the 3D terrain model

In the next step, we found the altitude for each vertex from the Daftlogic website [8]. On the basis of the obtained altitude values, we performed the modification of the spatial coordinate for each vertex. With this, we obtained the final shape of the hospital terrain model.

In addition, we created several simple models of trees and shrubs, which allowed us to make the 3D hospital area model more complex. These models were created by the deformation of the Sphere (treetops) and Cylinder (trunks) objects.

IV. Texturing, Animation and Rendering

We used the UV mapping technique for texturing objects. This process starts by the decomposition of each object into 2D sub-surfaces (a UV map). At the beginning of the decomposition process, it is necessary to mark the edges, which should be ripped from another one. In Blender, this process is performed by the Mark Seam command. After that, it is possible to finish decomposing by using the Unwrap tool. The UV map created in this way is saved into the .png raster graphic format (it is also possible to save it into another raster graphic format, but we need to use a lossless compression algorithm).

All textures were drawn in the GIMP software environment [9]. GIMP is a program under the GNU license. It is a freely distributed piece of software for such tasks as photo retouching, image composition and image authoring. It works on many operating systems and in many languages.

We also opened all UV maps in GIMP. In these pictures, the location of each part of the 3D object is visible. With this information, we can fill each individual sub-surface as necessary. Most of these textures were drawn by hand, and in some cases, we used pre-created textures from the CGTextures website [10] – these textures were edited and modified in order to use them on our models. For texture creation and editing purposes, we used standard GIMP drawing, coloring and transforming tools [4]. Once this process was finished, we saved all of the created textures back into same files and opened and mapped these on the appropriate 3D models in the Blender environment. An example of the drawn texture is shown in Figure 5.

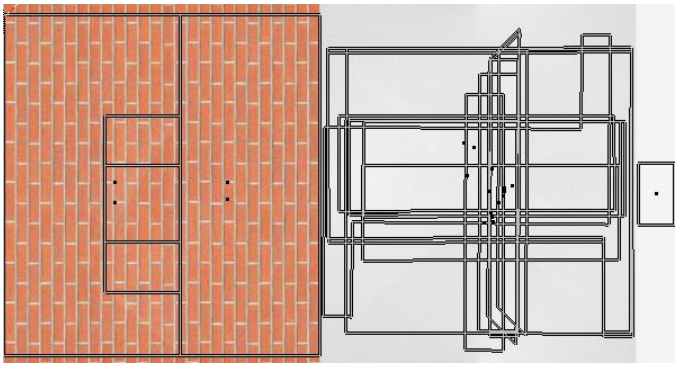


Fig. 5 the uv texture of the Baby box building

We also created and textured all 3D sub-models of each period in this way. In the end, we created a complex 3D scene, where we imported all of the created models and placed them in the correct places. We used this process for the scenes with the models of the hospital complex from 1927, 1930, 1935, 1940, 1950, 1960, 1980, 1990, 2000, 2005 and 2013.

Because we wanted the best visualization possible, we also created an animation for all complex scenes. These animations were performed by using the "bird´s-eye view". The technical solution of our animation was based on the animation curves that were followed by the rendering camera. Shaping the curve around the hospital complex and setting

the orientation of the camera allows us to obtain the desired visualization.

The Render command performs the rendering calculation process in the Blender environment. Additionally, we can set many of the accompanying parameters. The basic parameters are the choice of a rendering algorithm, image or animation resolution, type of output file format, antialiasing, motion blur, enable/disable ray-tracing and shadows. We made the decision to use Blender's internal renderer with an image resolution of 1280x720 pixels, 25 frames per second and the MPEG-2 output format to render animations. Figure 6 shows one frame of the rendered animation of the hospital area from 2000.

V. Conclusion

Our next goal is to expand and improve the current model, for example - to make the pavement models more detailed, to add the road on the edge of the Dřevnice River (this river is in the surrounding area of the hospital), more model types of trees, etc. The next improvement could be realized by the rendering of the current hospital area model in high resolution. After that, this render output could be stored on the hospital www pages and made into an interactive image, which could show information about each building, when the mouse cursor is placed on it. Another improvement is to create time animation, which visualizes the development of the hospital area over the years of its existence.

References

[1] J. Bakala, *Baťova nemocnice ve Zlíně 1927-2002*. Zlín, KODIAK print, s.r.o., 2002.

[2] J. Bakala, *Baťova nemocnice v obrazech, faktech a dokumentech*. Zlín, Finidr, s.r.o., 2002.

[3] R. Hess, *Blender Foundations – The essential Guide to Learning Blender 2.6*. Focal Press, 2010.

[4] A. Peck, *Beginning GIMP:From Novice to Professional*. Apress, 2008.

[5] Soka Zlín. (2012). Státní okresní archív Zlín, Moravský zemský archív v Brně [Online]. Available: http://www.mza.cz/zlin/

[6] SKILL production. (2014). starý Zlín - historie Zlína, pohlednice a video [Online]. Available: http://staryzlin.cz/

[7] Blender contributors. (2014). blender.org - Home [Online]. Available: http://www.blender.org

[8] DAFT Logic. (2014). Google Maps Find Altitude [Online]. Available: http://www.daftlogic.com/sandbox-google-maps-find-altitude.htm

[9] Gimp contributors. (2014). GIMP - The GNU Image Manipulation Program [Online]. Available: http://www.gimp.org

[10] CGTextures contributors. (2014). CG Textures – Textures for 3D, graphic design and Photoshop! [Online]. Available: http://www.cgtextures.com/

Fig. 6 one frame of the rendered animation of the hospital area from 2000

Possibility of Chest X-Ray Images for Image Guided Lung Biopsy System

Q. Rizqie, D.E.O Dewi, M. A. Ayob, I. Maolana, R. Hermawan, R. D. Soetikno, and E. Supriyanto

***Abstract*—**Biopsy is a diagnosis technique to detect cancerous cell by removal of sample cell from inside of the body . Advancement in imaging system give way to more precise biopsy. Nowadays, Computed Tomography (CT) is one of the most used imaging modalities for image guided biopsy, however, not every clinics and hospitals have access to CT machine especially in developing countries. This paper is highlight the possibility to use conventional radiography (X-Ray) as altenative imaging modalities for image guided lung biopsy. X-Ray is more available, have lower cost, and lower radiation level. Two X-Ray images from anterior-posterior position and lateral position is taken to compensate with X-Ray's two dimensional nature. Physician will mark the target in each images, then the system will put the images in three dimensional plot.

***Keywords*—**Computer Vision, Image Guided Systems , Medical Imaging, X-Ray

I. Introduction

WHO mention that Cancer is a leading cause of death worldwide, accounting for 7.6 million deaths (around 13% of all deaths) in 2008. Among them, lung cancer contributed around 1.37 million deaths[9].

Lung cancer symptoms usually take years to emerge, and only emerging while the cancer already advanced, because of this, some of physicians recommend people with high risk of lung cancer (i.e. people who smoke a lot) to do an early screening in order to check whether they have lung cancer or not. If screening results show something suspicious, such as an unidentified nodule inside lung, physicians will recommend biopsy procedure[8].

This work was supported by Cardio Centre Flagship Grant, Universiti Teknologi Malaysia.

Q. Rizqie, D.E.O Dewi, M. A. Ayob and E. Supriyanto are with IJN-UTM Cardiovascular Engineering Centre, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia (Corresponding author phone: +6075558551-mail: eko@biomedical.utm.my).

I.Maolana, R. Hermawan and R. Soetikno are with Department of Radiology, Faculty of Medicine, University of Padjajaran, Bandung, Indonesia.

Biopsy is a medical diagnosis technique involving removal and examination of cells from various part of human body in order to detect presence or extend of disease. Lung biopsy is biopsy done in order to help diagnose lung disease, including lung cancer and tumor. There are 4 type of lung biopsy, bronchoscopic biopsy, needle biospy, open biopsy, and video assisted biopsy.

Advancement in imaging technique, allowed greater precision for targeting suspected lessions while avoid important tissue such as blood vessel, thus minimize the risk and improve the accuracy of diagnosis. Imaging modalities used to guide biopsy are CT, Fluoroscopy, CT-Fluoroscopy, PET, PET-CT, and USG, though MRI also used in some cases.

One of the most used imaging modalities for lung biopsy is CT, this imaging modality give relatively good results because it is capable of imaging most of lung structures, however CT also raise concern about danger of radiation, eventhough the level of radiation is deemed save, multiple and prolonged exposure could still be dangerous. CT machine also not always available, especially in small hospital in developing countries. Unfortunately, popular image guided biopsy systems still depend heavily on CT imaging.

Compared to CT, two dimensional conventional radiography widely known as X-Ray, has a much lower rate of radiation. For chest radiography, full chest CT will expose the patient to ± 8.0 mSv, while X-Ray only expose the patient to ± 0.02 mSv for Posterior Anterior Position and ± 0.04 mSv for Lateral position[2]. X-Ray also relatively more widespread compared to CT machine, even a small hospital in developing countries can have one X-Ray machine. It is possible to detect lung nodule only using X-Ray, however it depends on the position and mass of the nodule.

II. Method

A. Computed Tomography

Computed Tomography (CT) is the first method to view the inside of human body without bias from superposition of the objects inside. CT is the main imaging modality whenever situation require fast three dimensional imaging. CT can display images in two dimensional slice or three dimensional volume. CT also capable of multiplanar reformation, a technique to change the view of three dimensional volume into two dimensional planes relative to orthogonal axes[5][6].

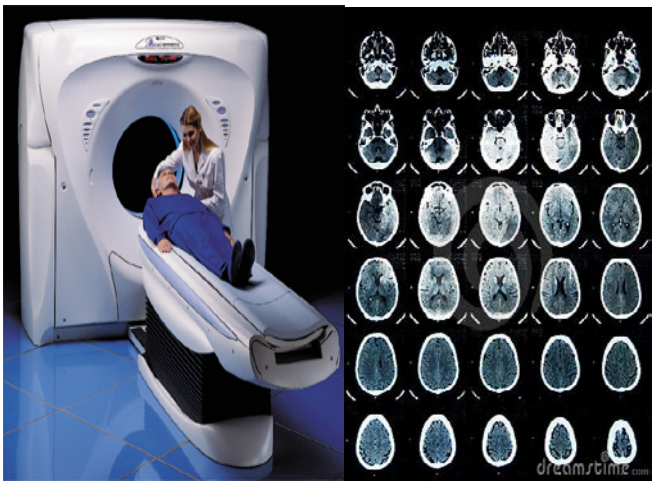


Fig. 1 CT Machine and CT Images

B. *Conventional Radiography (X-Ray)*

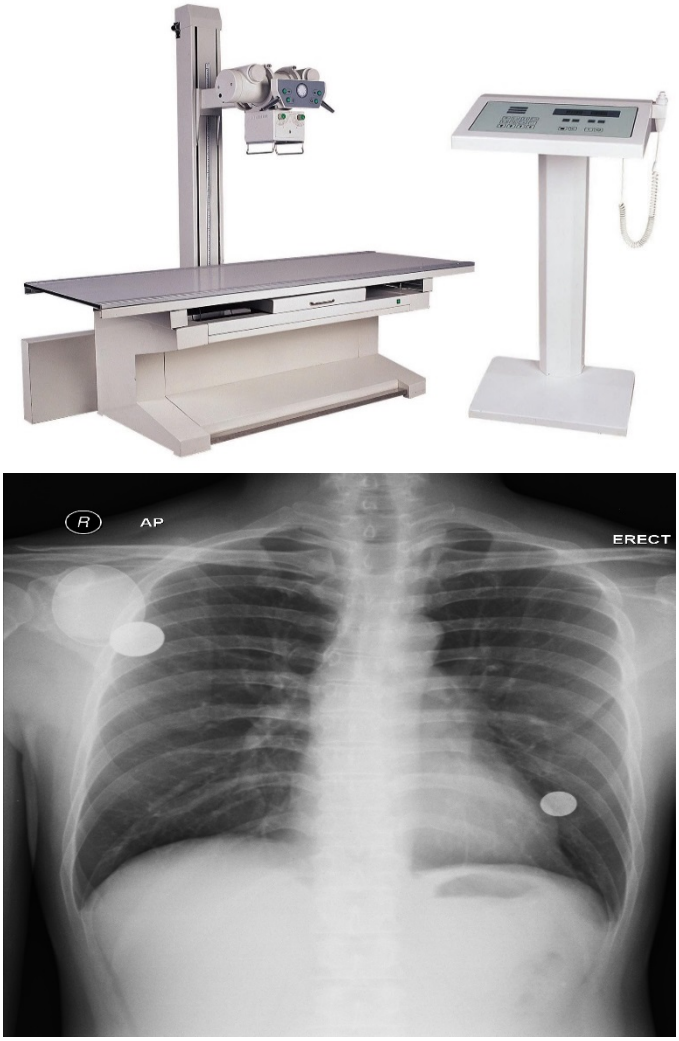


Fig. 2 X-Ray Machine and X-Ray Images

Conventional Radiography also known as roentgen, from Wilhem Roentgen, a scientist that discover the X-Rays used in radiography.Radiography is strongly associated with the term X-Ray, to the point if some one mentioned X-Ray, it almost always means radiography. X-Ray imaging can acquire large area of human body, such as complete chest, in one take, unfortunately, it's two dimensional nature, make the objects inside overlap each others[1].

C. *CT-Guided Biopsy*

CT-Guided biopsy procedure start with a CT image acquisition. Physician then will detect the position of targeted nodules.Biopsy procedure should be executed without delay after CT-Image acquisition. In case there are needs of pre-operative meeting, another CT acquisition must be done before biopsy execution, this is because there are possibilites that the targeted nodule could grow.Like any other intervention procedure, physician should wear protective clothes, and gloves[3].

Based on CT images acquired, physician will decide the best point to insert needle, thus patient will be ask to stay in the position that make it easier for physician to reach that point[3].

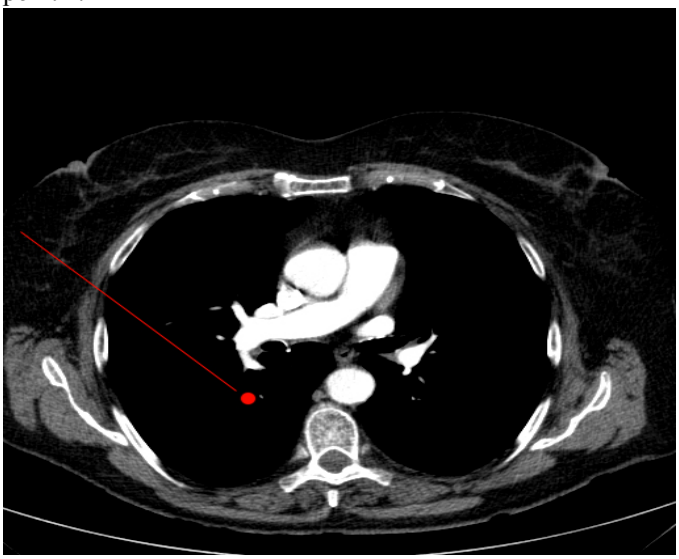


Fig. 3 Ilustration of needle path in CT image

Physician will find the most suitable needle entry point by analyze data from acquired CT Images. Physician will mark the target and others important feature, such as pulmonary blood vessel, which must be avoided by the needle. The skin entry point should be sterilised, a local anesthesia can also be delivered. During insertion of the needle patient will be asked to control their breathing, the technique to do this should be teached and practised before needle insertion.Patient should completely suspend their breathing when the needle is advanced or withdrew[3][7].

D. *Proposed System*

The use of chest X-Ray images, instead of CT images as media for deciding the entry point and target is proposed. Two X-Ray images will be used, one is posterior anterior of chest X-Ray, and the other are from lateral left and/or right of chest X-Ray.

ISBN: 978-1-61804-236-1

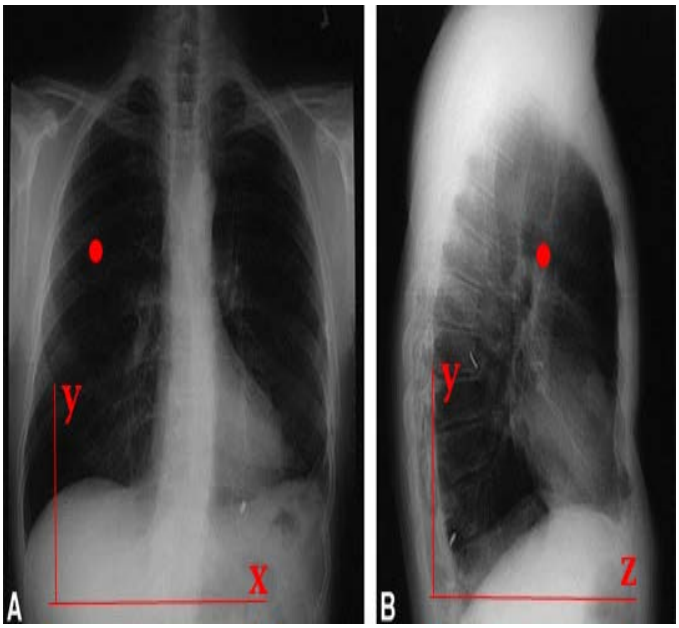


Fig.4 Ilustration how to find biopsy target in X-Ray image

Physician can pick the targeted point in each images, the system then will registrate the targeted points in both images and visualize them in 3D plot.

Simple formula to registrate each point to suitable point in 3D Plot, is as follow.

$$\alpha = \frac{x1}{height} \quad ; \quad \beta = \frac{(y1 - y2)}{width} \quad ; \quad \gamma = \frac{x2}{height}$$

$$x' = 0.5 + \alpha \; ; y' = \left(\frac{y}{height}\right) - 0.5 + \beta \; ; \; z' = \left(\frac{x}{width}\right) - \gamma$$

III. Experiment

Protoype of the system have been developed, this prototype lack the capability to accurately guide physician to execute biopsy procedure, however this prototype shows the possibility of conventional X-Ray to be used as alternative of CT guided lung biopsy.

This prototype developed under assumption x-ray images have DICOM format which including the pixel spacing information that needed to transform pixels distance into real world distance.

This prototype developed using GDCM library to read the dicom, OpenCV library for pre-processing, and OpenGL library for visualization.

Simple experiment have been done in order to test the protoype system. X-ray images of a person is acquired, a small coin is attached to that person back. From posterior anterior view image, the coin looks like a bright circle, while from lateral view of image, it looks like an elipse object. User of the system will manually click the object of interest in both images. The system then put both images in three dimensonal point of view. By crossing point of interest in both images, the position of the target can be seen in three dimensional view.

Flowchart of the prototype is as follow :

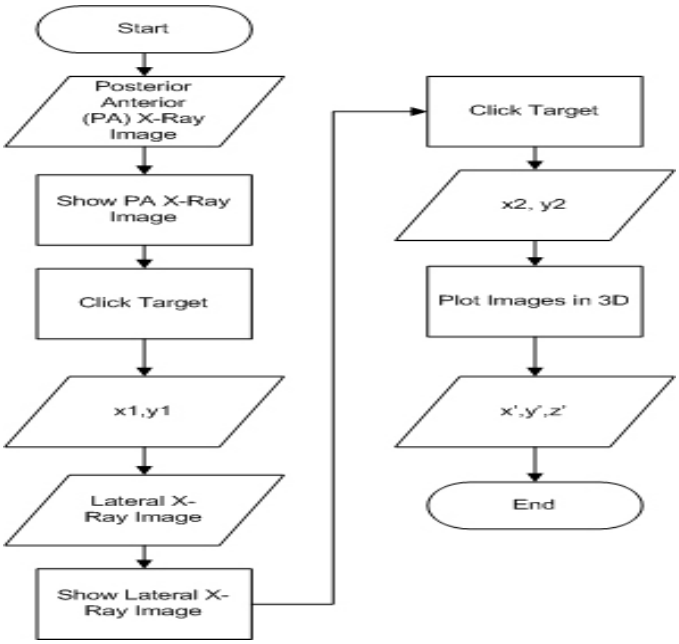
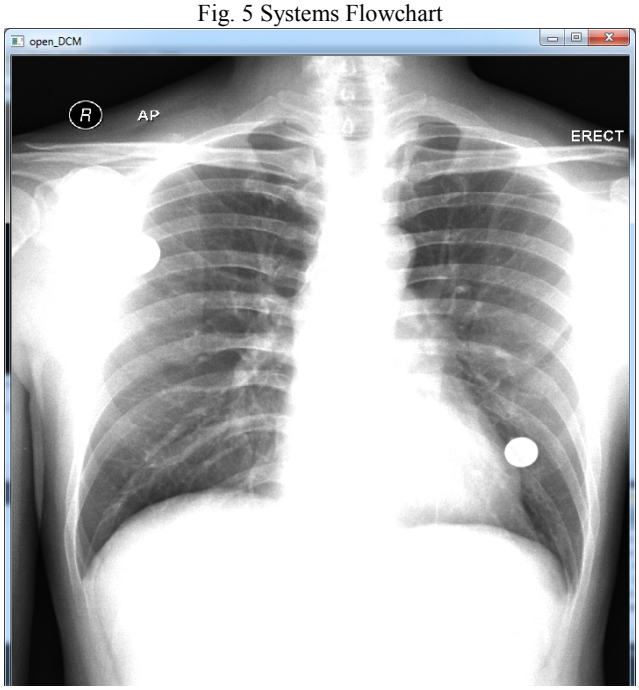


Fig. 5 Systems Flowchart



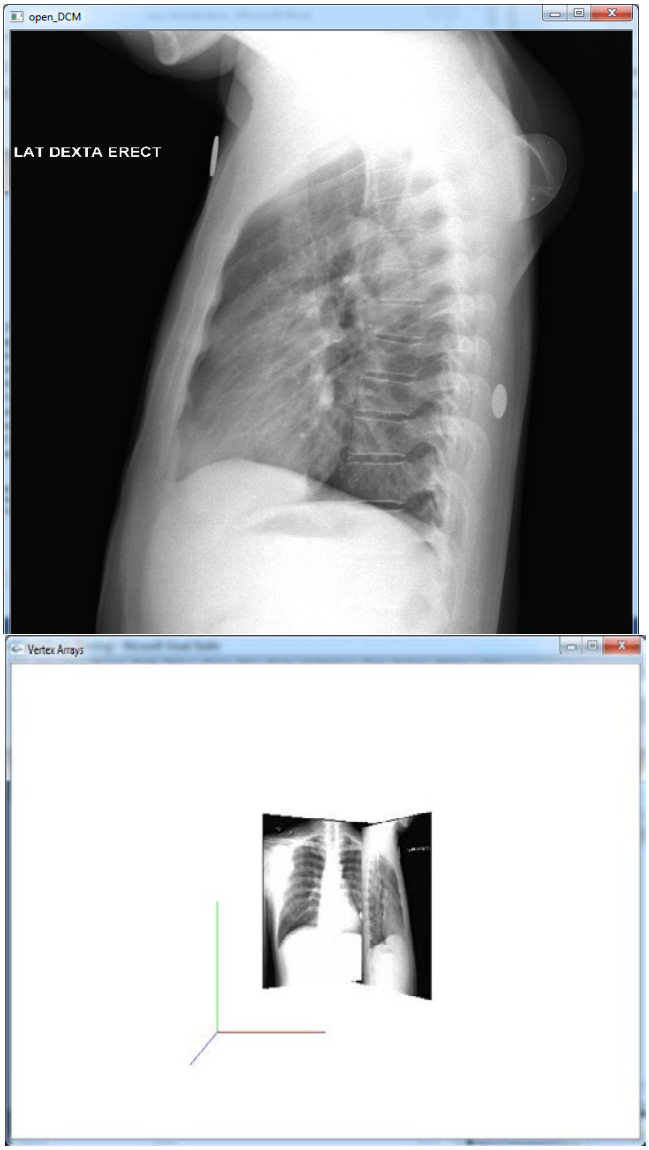


Fig. 6 Systems Output

IV. Future Works

It is clear from the experiment that CT guided biopsy is easier to do. While X-Ray have capabilities to be used as an alternative, and canpoint the target, the system still need more mesurement than CT guided system. It is imperative to develop the system so physician will have the same comfort with using CT guided system. The prototype system also cannot recognize ribs and pulmonary vessel, while ini CT guided system, ribs and pulmonary vessels easily pointed, and physician can plan a needle path that avoid both of ribs and pulmonary vessels. The system developed should address this issue and have capabilities to shows important feature such as ribs and pulmonary vessels clearly.

.

Acknowledgment

Author like to thanks radiology division of Hasan Sadikin Hospital and radiology department of medical faculty of universitas padjajaran for aids in acquisition of image used in experiment.

References

[1] SS.Hiss, *Understanding Radiography*, Charles C. Thomas Publisher, Limited, 1993.
[2] BF.Wall, D.Hart. *Revised radiation doses for typical x-ray examinations.* The British Journal of Radiology 70:437-439; 1997
[3] A .Manhire, M.Charig, C.Clelland, et al. *Guidelines for radiologically guided lung biopsy*. Thorax 2003;58:920-36
[4] JK.Gohagan, PM.Marcus, RM.Fagerstrom, et al. *Final results of the Lung Screening Study, a randomized feasibility study of spiral CT versus chest x-ray screening for lung cancer*. Lung Cancer. 2005;47(1):9–15.
[5] TM.Buzug, *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*, Springer, 2008.
[6] J.Hsieh, *Computed Tomography : Principles, Design, Arrtifacts, and Recent Advances*, Wiley, 2009.
[7] I-Chen Tsai, Wei-Lin Tsai, Min-Chi Chen, Gee-Chen Chang, Wen-Sheng Tzeng, Si-Wa Chan, and Clayton Chi-Chang Chen, *CT-Guided Core Biopsy of Lung Lesions: A Primer*, American Journal of Roentgenology 2009 193:5, 1228-1235.
[8] www.makna.org.my
[9] www.who.int

A Heuristic Cluster-Head Selection Algorithm for Clustering-Based Wireless Sensor Networks: Based on VIKOR Technique

Hossein Jadidoleslamy

Information Technology Engineering Group, Department of Information Technology, Communications and Security, Malekashtar University of Technology (MUT), Tehran, Iran
E-mail: Tanha.hossein@gmail.com

***Abstract*—**Wireless Sensor Networks (WSNs) are consisting of many sensor nodes and a Sink. Some problems of WSNs are their hard management, limited resources, dynamic topology and low scalability. One significant solution against these problems is clustering. In clustering process, Cluster-Head (CH) selection is an important step, especially in clustering-based and homogenous WSNs. As a result, this paper is represented a centralized, multi-criteria and a top-down CH selection algorithm. The proposed algorithm is based on the VIKOR technique; it is ranking members of each cluster based on several non-commensurable and conflicting criteria and then, selecting appropriate CH for each cluster.

***Keywords*—**Wireless Sensor Network (WSN), Clustering, Cluster-Head Selection, VIKOR, Multi-Criteria, Energy-Aware.

I. INTRODUCTION

Wireless Sensor Networks (WSNs) are homogeneous or heterogeneous networks which they consist of many sensor nodes, some Cluster-Heads (CHs) and a Sink. CHs collect data from their nearby sensor nodes; then, aggregating and forwarding them to the Sink for final processes [1, 2]. Some of most important problems in WSNs are hard organization and management, limited resources, dynamic topology and low scalability; but, clustering is an effective solution to solving these problems. So, clustering is a vital and complex requirement for WSNs; since, it leads to more scalability, energy efficiency, prolonged network lifetime and manageability of large-scale WSNs. As a result, this paper will be discussed about WSNs' clustering. The main purpose of this paper is focusing on CH selection step of the clustering process; due to CH selection is an important step, especially in clustering-based and homogenous WSNs. Thus, this paper is represented a centralized, multi-criteria and a top-down CH selection algorithm for clustering-based and homogenous WSNs. It is using of different non-commensurable and conflicting criteria for CH selection such as sensor nodes' Remainder Energy (RE), Distances to the clusters' Gravity Center (DGC), Distance Average between each node and other members of the cluster (DA), Distance to the Sink (DS) and Distance to the node with Minimum Remainder Energy into the cluster (DMRE). Also, the proposed algorithm is based on VIKOR technique; it is ranking members of each cluster and then, selecting appropriate CH for each cluster. All of calculations will be done by the Sink through using of mathematical functions and quantitative variables, in centralized. Therefore, the proposed technique is a top-down and centralized CH selection approach.

The VIKOR technique [3, 4] introduced the multi-criteria ranking index based on the particular measure of closeness to the ideal solution and it was introduced as one applicable technique to implement within the MCDM (Multi-Criteria Decision Making) approach [4]. The VIKOR method was developed as a multi-criteria decision making method to solving discrete decision making problems with several criteria. This method focuses on ranking alternatives and selecting appropriate alternative in the presence of several non-commensurable and conflicting criteria, which could help the decision makers to reach a final decision. The MCDM approach consists of constructing a global preference relation for a set of alternatives evaluated using several non-commensurable and conflicting criteria; then, selecting the best alternative [4, 5, 6].

Rest of this paper is organized as following: Section2 expressed an overview of WSNs; Section3 is considering WSNs' clustering; Section4 is represented the proposed CH selection algorithm; Section5 expressed a case study (a practical experiment); and finally, Section6 is presented conclusion and future works.

II. AN OVERVIEW OF WIRELESS SENSOR NETWORKS

Wireless Sensor Network (WSN) is a wireless computer network with following main features [1, 2, 7, 8]:

- Infrastructure-less;
- No public address (data-centric);
- Consists of many tiny sensor nodes (small size, low-cost and low-power);
- High density nodes distribution in operational environment;
- Insecure radio links;
- Different architectures: (hierarchical or flat), (centralized or distributed) and (homogenous or heterogeneous);
- Limited resources (radio range, bandwidth, energy, memory and processing);
- Main application domains: monitoring and tracking;

In continue, it is represented different aspects of WSNs, such as characteristics, applications, communication architectures and vulnerabilities (as Figure1).

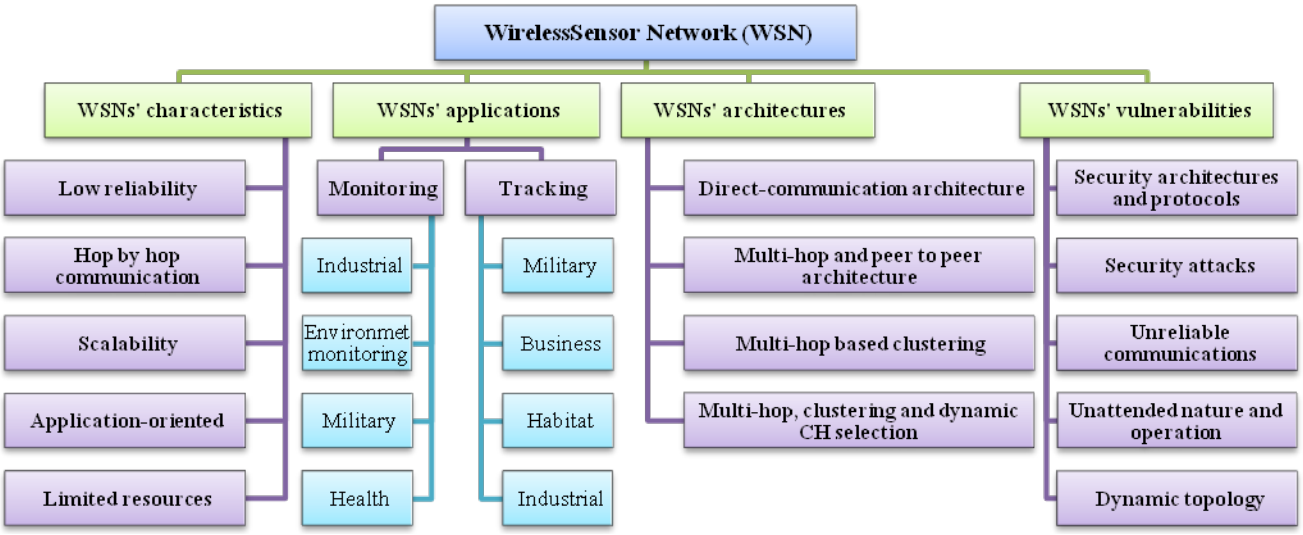


Fig. 1. Different dimensions of WSNs

III. CLUSTERING IN WIRELESS SENSOR NETWORKS

According to the Figure2, clustering means dividing the WSN's nodes in virtual groups (called clusters). In other words, clustering is grouping nodes (clusters formation) and selecting a Cluster-Head (CH) for each group. Each CH gathers data from its associated nodes and forwarding the aggregated data to the Sink [7, 9, 10, 11].

Thus, clustering has three main phases, including of: clusters formation, CHs selection and steady state phase [12, 13, 14]. In continue Figure3 represents some common goals, necessity and advantages of WSNs' clustering [15, 16]. Also, Figure4 is representing different classifications of clustering algorithms in WSNs [7, 16, 17, 18].

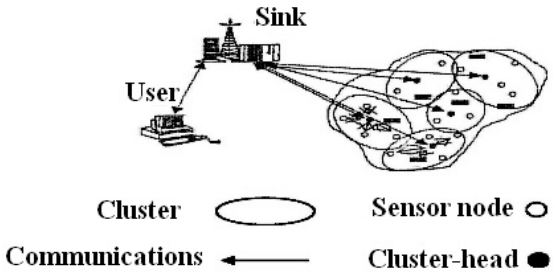


Fig. 2. An overview of clustering in WSNs

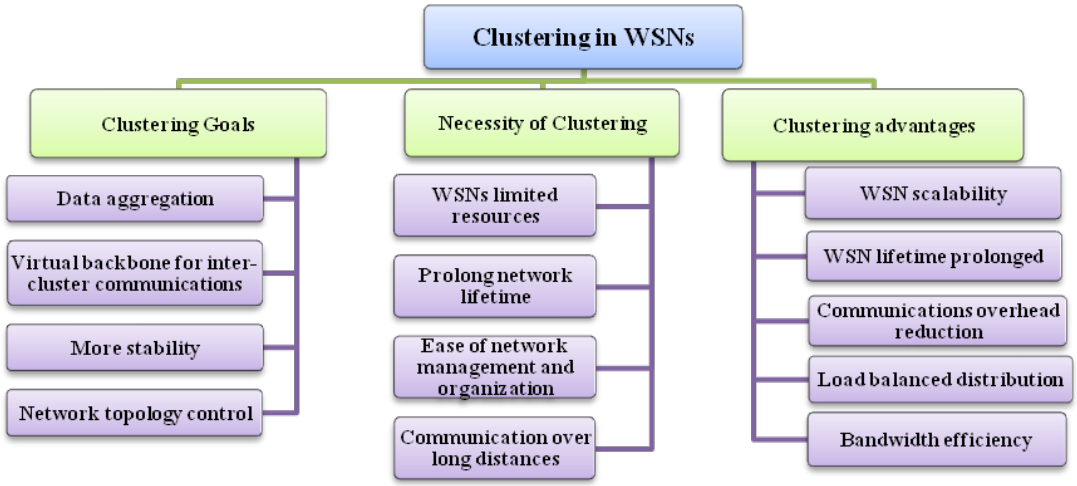


Fig. 3. Goals, necessity and advantages of clustering in WSNs

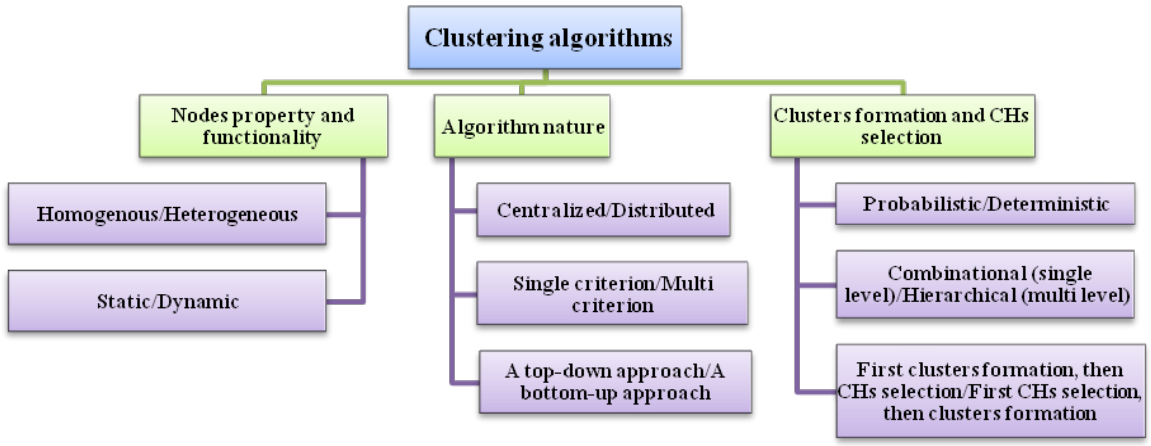


Fig. 4. Different classifications of clustering algorithms in WSNs

IV. The Proposed Cluster-Head Selection Algorithm: Based on the VIKOR Technique

In each cluster with n nodes and m criteria, steps of the proposed algorithm are as following:

Step1: Determining ranking and selection criteria; then, constructing the decision making matrix (A); it is a n*m matrix, as:

- n= count of nodes;
- m= count of criteria;
- Let x_{ij} is score of node i with respect to criterion j;
- Ranking and selection criteria are including of: positive criteria (more is better) and negative criteria (less is better);

By attention to the evaluation of nodes for different criteria, the decision making matrix is constructing as following:

A=

Node/Criterion	Criterion 1	Criterion 2	…	…	Criterion m
Node 1	A_{11}	A_{12}	…	…	A_{1m}
Node 2	A_{21}	A_{22}	…	…	A_{2m}
…..	…	…	…	…	…
…..	…	…	…	…	…
Node n	A_{n1}	…	…	…	A_{nm}

Note: A_{ij}= functionality of node i'th relative to criterion j'th;

Step2: Determining vector of criteria's weight: in this step, by attention to the importance factor of different criteria in decision making process, the weight vector will be defined.

- $W = \{W_1, W_2, \ldots, W_m\}$, as $\sum_{j=1}^{m} (W_1 + W_2 + \cdots + W_m) = 1$;

Step3: Specifying the positive ideal and negative ideal point: for each criterion, the best value and the worst value (between all nodes) will be determined.

- Best value for criterion j'th = f_{j*}
- Worst value for criterion j'th = f_{j-}

Step4: Calculating value of earning (S) and value of regret (R) for each node as following:

- S_i: relative distance of node i'th than positive ideal solution (the best combination);

$$S_i = \sum_{j=1}^{m} W_j \times \frac{(f_{j^*} - f_{ij})}{(f_{j^*} - f_{j-})}$$

- R_i: maximum regret of node i'th than farness of positive ideal solution;

$$R_i = \text{Max}\ \{W_j \times \frac{(f_{j^*} - f_{ij})}{(f_{j^*} - f_{j-})}\};$$

Step5: Calculating the VIKOR index (Q) for each node by using of the following formula:

- $Q_i = V \times \frac{(S_i - S^*)}{(S^- - S^*)} + (1 - V) \times \frac{(R_i - R^*)}{(R^- - R^*)}$
- $S^- = \text{Max}\ S_i$, $S^* = \text{Min}\ S_i$
- $R^- = \text{Max}\ R_i$, $R^* = \text{Min}\ R_i$
- $V \in [0, 1]$;

Step6: Ranking nodes based on Q, S and R values in descending order. Node A' which it's Q is minimum, it is an appropriate solution; if A' has been following conditions, it will be ranked as the best node.

- **Condition (1):** acceptable advantage:

$Q(A'') - Q(A') \geq$, n: count of nodes, A": second node into the Q ranking list;

- **Condition (2):** acceptable stable in decision making: Node A' also should be had highest rank in the S or R or both ranking lists.

Note: if one of above conditions be not established, a set of solutions will be proposed, as following:

- The set of proposed solutions = {A', A"}, if only the second condition be not established;
- The set of proposed solutions = {A', A", … , A^k}, if the first condition be not established; A_k be determined by using of following relation for the maximum value of k:

$Q(A^k) - Q(A') <$;

Note: For each cluster, time will be divided into some super-round; each super-round has some rounds; count of rounds of each super-round is equal to count of nodes of that cluster. The proposed algorithm is ranking and prioritizing the cluster's members for being CH in different rounds of the super-round, in order (at the first of each super-round).

Note: If the count of a cluster's members be greater than a predefined threshold, it is better to breaking that cluster to some new smaller clusters (with less count of nodes).

V. PRACTICAL EXPERIMENT: A CASE STUDY

This section wants to prioritizing and ranking nodes of each cluster of following WSN (Figure5) to selecting the appropriate CHs. Our assumptions are:

- Operational environment dimensions = 6 m * 4 m;
- Count of all sensor nodes: N = 9; Count of CHs or Clusters: K = 2;
- Ranking and selection criteria are as following:
  - Criterion1: Remainder Energy: RE
  - Criterion 2: Distance to the Gravity Center: DGC
  - Criterion3: Distance Average between each node and other members of the cluster: DA
  - Criterion4: Distance to the Sink: DS
  - Criterion5: Distance to the node with Minimum Remainder Energy into the cluster: DMRE

  Therefore, ranking and selection criteria are: {RE, DGC, DA, DS, DMRE}
- Positive utility Criterion: Earnings Criterion = {RE}
- Negative utility Criterion: Cost Criterion = {DGC, DA, DS, DMRE}
- The Sink deployment location coordinates = gravity center of the WSN = (3.66, 2.77)
- Cluster1 is including of {node1, node2, node4, node5, node6, node7};
- Cluster2 is including of {node3, node8, node9};
- Gravity center coordinates of cluster1:

Length of the cluater1's gravity center coordinates = $\underline{(3+4+4+3+2+2)}$

$$\frac{\sum_{i=1}^{6} Xi}{6} = \frac{}{6} = \frac{18}{6} = 3$$

Width of the cluster1's gravity center coordinates = $\underline{(4+4+3+3+4+3)}$

$$\frac{\sum_{i=1}^{6} Yi}{6} = \frac{}{6} = \frac{21}{6} = 3.5$$

⇨ Coordinates of cluster1's gravity center = (3, 3.5);

- Gravity center coordinates of cluster2:

Length of the cluater2's gravity center coordinates = $\underline{(5+4+6)}$

$$\frac{\sum_{i=1}^{3} Xi}{3} = \frac{}{3} = \frac{15}{3} = 5$$

Width of the cluster2's gravity center coordinates = $\underline{(1+1+2)}$

$$\frac{\sum_{i=1}^{3} Yi}{3} = \frac{}{3} = = 1.33$$

⇨ Coordinates of cluster2's gravity center = (5, 1.33);

- Sensor nodes are distributed in operational environment, randomly; supposing they are deployed on following coordinates (according to the Figure5);

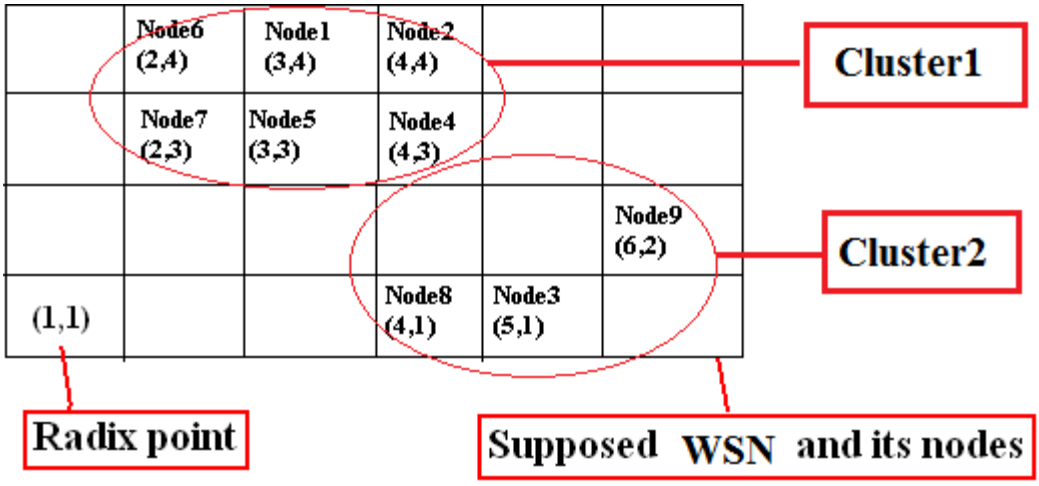


Fig. 5. Case study (practical experiment environment)

TABLE 1. VALUES OF ALL NODES' RANKING AND SELECTION CRITERIA FOR CHS SELECTION

Sensor node	Deployment location coordinates	RE	DGC	DA	DS	DMRE
1	(3, 4)	10	0.50	0.97	$\sqrt{0.6^2+1.2^2}=1.4$	1.41
2	(4, 4)	8	1.12	1.28	$\sqrt{0.3^2+1.2^2}=1.2$	1
3	(5, 1)	3	0.33	0.80	$\sqrt{1.3^2+1.7^2}=2.2$	0
4	(4, 3)	5	1.12	1.28	$\sqrt{0.3^2+0.2^2}=0.4$	0
5	(3, 3)	11	0.50	0.97	$\sqrt{0.6^2+0.2^2}=0.7$	1
6	(2, 4)	7	1.12	1.28	$\sqrt{1.6^2+1.2^2}=2.0$	2.24
7	(2, 3)	7	1.12	1.28	$\sqrt{1.6^2+0.2^2}=1.6$	2
8	(4, 1)	8	1.05	1.08	$\sqrt{0.3^2+1.7^2}=1.8$	1
9	(6, 2)	7	1.20	1.22	$\sqrt{2.3^2+0.7^2}=2.4$	1.41

A. Nodes' ranking and Cluster-Head selection for Cluster1: Determining CH1

Step1: The decision making matrix (A) for cluster1 is as following:

A =

Sensor node/Criteria	RE	DGC	DA	DS	DMRE
Node1	10	0.50	0.97	1.40	1.41
Node2			1.28	1.28	
Node4			1.28	0.41	
Node5	11	0.50	0.97	0.70	
Node6			1.28	2.07	2.24
Node7			1.28	1.68	

Step2: Determining the weight vector by attention to the importance of each criterion;

$W = \{W_1, W_2, W_3, W_4, W_5\} = \{0.2, 0.2, 0.2, 0.2, 0.2\}$

Note: effective percentage of each criterion = 20 % = 0.2; if all criteria (i.e. RE, DGC, DA, DS and DMRE) have been equal effective percentage (importance), their weight is equal to each other (i.e. 0.2); else, they have different weights as: $\sum_{i=1}^{5} Wi = 1$;

Step3: Determining best and worst values for different criteria in cluster1: the ideal solution in positive criteria (i.e. RE) is the maximum value into the related column and in negative criteria (i.e. DGC, DA, DS and DMRE) is the minimum value into the related column.

Note: Best value = f_{j*}, Worst value = f_{j-} ;

Values	C_1	C_2	C_3	C_4	C_5
f_{j*}	11	0.50	0.97	0.41	0
f_{j-}	5	1.12	1.28	2.07	2.24

Step4: Determining value of earning criteria (S) and value of regret criteria (R) for each node, as following:

$S_1 = 0.2 * \frac{11-10}{11-5} + 0.2 * \frac{0.50-0.50}{0.50-1.12} + 0.2 * \frac{0.97-0.97}{0.97-1.28} + 0.2 * \frac{0.41-1.40}{0.41-2.07} + 0.2 * \frac{0-1.41}{0-2.24} = 0.278$

$R_1 = \text{Max } \{0.033, 0, 0, 0.119, 0.126\} = 0.126$

$S_2 = 0.2 * \frac{11-8}{11-5} + 0.2 * \frac{0.50-1.12}{0.50-1.12} + 0.2 * \frac{0.97-1.28}{0.97-1.28} + 0.2 * \frac{0.41-1.28}{0.41-2.07} + 0.2 * \frac{0-1}{0-2.24} = 0.694$

$R_2 = \text{Max } \{0.1, 0.2, 0.2, 0.105, 0.089\} = 0.2$

$S_4 = 0.2 * \frac{11-5}{11-5} + 0.2 * \frac{0.50-1.12}{0.50-1.12} + 0.2 * \frac{0.97-1.28}{0.97-1.28} + 0.2 * \frac{0.41-0.41}{0.41-2.07} + 0.2 * \frac{0-0}{0-2.24} = 0.6$

$R_4 = \text{Max } \{0.2, 0.2, 0.2, 0, 0\} = 0.2$

$S_5 = 0.2 * \frac{11-11}{11-5} + 0.2 * \frac{0.50-0.50}{0.50-1.12} + 0.2 * \frac{0.97-0.97}{0.97-1.28} + 0.2 * \frac{0.41-0.70}{0.41-2.07} + 0.2 * \frac{0-1}{0-2.24} = 0.124$

$R_5 = \text{Max } \{0, 0, 0, 0.035, 0.089\} = 0.089$

$S_6 = 0.2 * \frac{11-7}{11-5} + 0.2 * \frac{0.50-1.12}{0.50-1.12} + 0.2 * \frac{0.97-1.28}{0.97-1.28} + 0.2 * \frac{0.41-2.07}{0.41-2.07} + 0.2 * \frac{0-2.24}{0-2.24} = 0.933$

$R_6 = \text{Max } \{0.133, 0.2, 0.2, 0.2, 0.2\} = 0.2$

$S_7 = 0.2 * \frac{11-7}{11-5} + 0.2 * \frac{0.50-1.12}{0.50-1.12} + 0.2 * \frac{0.97-1.28}{0.97-1.28} + 0.2 * \frac{0.41-1.68}{0.41-2.07} + 0.2 * \frac{0-2}{0-2.24} = 0.865$

$R_7 = \text{Max } \{0.133, 0.2, 0.2, 0.153, 0.179\} = 0.2$

Sensor node	S	R
Node1	0.278	0.126
Node2	0.694	0.2
Node4	0.6	0.2
Node5	0.124	0.089
Node6	0.933	0.2
Node7	0.865	0.2

Step5: Now, calculating the VIKOR index value (as V = 0.5), for different nodes as following:

$S^- = \text{Max } \{S_1, S_2, S_4, S_5, S_6, S_7\} = \{0.278, 0.694, 0.6, 0.124, 0.933, 0.865\} = 0.933$

$S^* = \text{Min } \{S_1, S_2, S_4, S_5, S_6, S_7\} = \{0.278, 0.694, 0.6, 0.124, 0.933, 0.865\} = 0.124$

$R^- = \text{Max } \{R_1, R_2, R_4, R_5, R_6, R_7\} = \{0.126, 0.2, 0.2, 0.089, 0.2, 0.2\} = 0.2$

$R^* = \text{Min } \{R_1, R_2, R_4, R_5, R_6, R_7\} = \{0.126, 0.2, 0.2, 0.089, 0.2, 0.2\} = 0.089$

$Q_1 = 0.5 * \frac{0.278-0.124}{0.933-0.124} + 0.5 * \frac{0.126-0.089}{0.2-0.089} = 0.095+0.167 = 0.262$

$Q_2 = 0.5 * \frac{0.694-0.124}{0.933-0.124} + 0.5 * \frac{0.2-0.089}{0.2-0.089} = 0.352+0.5 = 0.852$

$Q_4 = 0.5 * \frac{0.6-0.124}{0.933-0.124} + 0.5 * \frac{0.2-0.089}{0.2-0.089} = 0.294+0.5 = 0.794$

$Q_5 = 0.5 * \frac{0.124-0.124}{0.933-0.124} + 0.5 * \frac{0.089-0.089}{0.2-0.089} = 0$

$Q_6 = 0.5 * \frac{0.933-0.124}{0.933-0.124} + 0.5 * \frac{0.2-0.089}{0.2-0.089} = 1$

$Q_7 = 0.5 * \frac{0.865-0.124}{0.933-0.124} + 0.5 * \frac{0.2-0.089}{0.2-0.089} = 0.458+0.5 = 0.958$

Step6: Ranking the nodes based on descending order; so, the nodes' ranking will be as following:

Q		S		R		Ranking
Node6	1	**Node6**	0.933	**Node6**	0.2	6
Node7	0.958	**Node7**	0.865	**Node7**	0.2	5
Node2	0.852	**Node2**	0.694	**Node2**	0.2	4
Node4	0.794	**Node4**	0.6	**Node4**	0.2	3
Node1	0.262	**Node1**	0.278	**Node1**	0.126	2
Node5	0	**Node5**	0.124	**Node5**	0.089	1

$Q_1 - Q_5 = 0.262 - 0 = 0.262$ and ≥ 0.2

Condition (1): $0.262 \geq 0.2$ ⇨ Acceptable

Condition (2): Node5 has highest rank into the S and R ranking list.

Order of being CH in Cluster1: Node5 ⇨ Node1 ⇨ Node4 ⇨ Node2 ⇨ Node7 ⇨ Node6

Best node or first CH for Cluster1: Node5

Worst node: Node6

⇨ As a result, the proposed algorithm selects Node5, Node1, Node4, Node2, Node7 and Node6 as CH1 candidate nodes, in order. In other words, in this super-round, sensor nodes will be selected as CH1 according to the following order:

Round1: Node5; Round2: Node1;
Round3: Node4; Round4: Node2;
Round5: Node7; Round6: Node6;

B. Nodes' ranking and Cluster-Head selection for Cluster2: Determining CH2

Step1: The decision making matrix (A) for cluster2 is as following:

A =

Sensor node/Criteria	RE	DGC	DA	DS	DMRE
Node3			0.8U	2.2Z	0
Node8			1.0U	1.8U	
Node9			1.2Z	2.46	1.41

Step2: Determining the weight vector by attention to the importance of each criterion;

$W = \{W_3, W_8, W_9\} = \{0.2, 0.2, 0.2\}$

Note: effective percentage of each criterion = 20 % = 0.2; if all criteria (i.e. RE, DGC, DA, DS and DMRE) have been equal effective percentage (importance), their weight is equal to each other (i.e. 0.2); else, they have different weights as: $\sum_{i=1}^{5} Wi = 1$;

Step3: Determining best and worst values for different criteria in cluster2: the ideal solution in positive criteria (i.e. RE) is the maximum value into the related column

and in negative criteria (i.e. DGC, DA, DS and DMRE) is the minimum value into the related column.
Note: Best value = f_{j*}, Worst value = f_{j-} ;

Values	C_1	C_2	C_3	C_4	C_5
f_{j*}	8	0.33	0.80	1.80	0
f_{j-}	3	1.20	1.22	2.46	1.41

Step4: Determining value of earning criterion (S) and value of regret criterion (R) for each node, as following:

$S_3 = 0.2 * \frac{8-3}{8-3} + 0.2 * \frac{0.33-0.33}{0.33-1.20} + 0.2 * \frac{0.80-0.80}{0.80-1.22} + 0.2 * \frac{1.80-2.22}{1.80-2.46} + 0.2 * \frac{0-0}{0-1.41} = 0.327$
$R_3 = \text{Max } \{0.2, 0, 0, 0.127, 0\} = 0.2$

$S_8 = 0.2 * \frac{8-8}{8-3} + 0.2 * \frac{0.33-1.05}{0.33-1.20} + 0.2 * \frac{0.80-1.08}{0.80-1.22} + 0.2 * \frac{1.80-1.80}{1.80-2.46} + 0.2 * \frac{0-1}{0-1.41} = 0.441$
$R_8 = \text{Max } \{0, 0.166, 0.133, 0, 0.142\} = 0.166$

$S_9 = 0.2 * \frac{8-7}{8-3} + 0.2 * \frac{0.33-1.20}{0.33-1.20} + 0.2 * \frac{0.80-1.22}{0.80-1.22} + 0.2 * \frac{1.80-2.46}{1.80-2.46} + 0.2 * \frac{0-1.41}{0-1.41} = 0.840$
$R_9 = \text{Max } \{0.04, 0.2, 0.2, 0.2, 0.2\} = 0.2$

Sensor node	S	R
Node3	0.327	0.2
Node8	0.441	0.166
Node9	0.840	0.2

Step5: Now, calculating the VIKOR index value (as V = 0.5), for different nodes as following:
$S^- = \text{Max } \{S_3, S_8, S_9\} = \{0.327, 0.441, 0.840\} = 0.840$
$S^* = \text{Min } \{S_3, S_8, S_9\} = \{0.327, 0.441, 0.840\} = 0.327$
$R^- = \text{Max } \{R_3, R_8, R_9\} = \{0.2, 0.166, 0.2\} = 0.2$
$R^* = \text{Min } \{R_3, R_8, R_9\} = \{0.2, 0.166, 0.2\} = 0.166$

$Q_3 = 0.5 * \frac{0.327-0.327}{0.840-0.327} + 0.5 * \frac{0.2-0.166}{0.2-0.166} = 0 + 0.5 = 0.5$

$Q_8 = 0.5 * \frac{0.441-0.327}{0.840-0.327} + 0.5 * \frac{0.166-0.166}{0.2-0.166} = 0.111 + 0 = 0.111$

$Q_9 = 0.5 * \frac{0.840-0.327}{0.840-0.327} + 0.5 * \frac{0.2-0.166}{0.2-0.166} = 0.5 + 0.5 = 1$

Step6: Ranking the nodes based on descending order; so, the nodes' ranking will be as following:

Q		S		R		Ranking
Node9	1	**Node9**	0.840	**Node9**	0.2	3
Node3	0.5	**Node8**	0.441	**Node3**	0.2	2
Node8	0.111	**Node3**	0.327	**Node8**	0.166	1

$Q_3 - Q_8 = 0.5 - 0.111 = 0.389$ and $= 0.5$
Condition (1): $0.389 \geq 0.5$ ⇨ Non acceptable ⇨ A set of solutions are:
$Q_3 - Q_8 = 0.389 < 0.5$
Then, Set of solutions: {Node8, Node3}

Order of being CH for Cluster2: Node8 ⇨ Node3 ⇨ Node9
Best node or the first CH for Cluster2: Node8
Worst node: Node9
⇨ As a result, the proposed algorithm selects Node8, Node3 and Node9 as CH2 candidate nodes, in order. In other words, in this super-round, sensor nodes will be selected as CH2 according to the following order:
Round1: Node8; Round2: Node3; Round3: Node9;

VI. Conclusion and Future Works

Nowadays, clustering in WSNs is a high interest area. This paper tried to discuss on main dimensions of WSNs and their clustering algorithms. Also, it represented centralized, multi-criteria and a top-down CH selection algorithm based on the VIKOR technique for clustering-based, homogenous and hierarchical WSNs. Figure6 is represented steps of the proposed algorithm and Figure7 is showing its different aspects and properties. This research has also included a practical case to show the effectiveness and feasibility of the proposed algorithm. The CH selection performed by using the proposed algorithm is able not only to find the appropriate CH to enhance the WSN's lifetime, but also to help the Sink to understand the ranking orders of the cluster's members to being CH in rounds of a super-round.

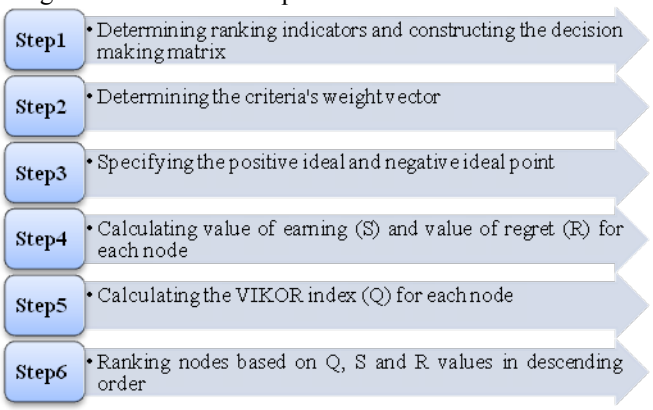


Fig. 6. Different steps of the proposed algorithm

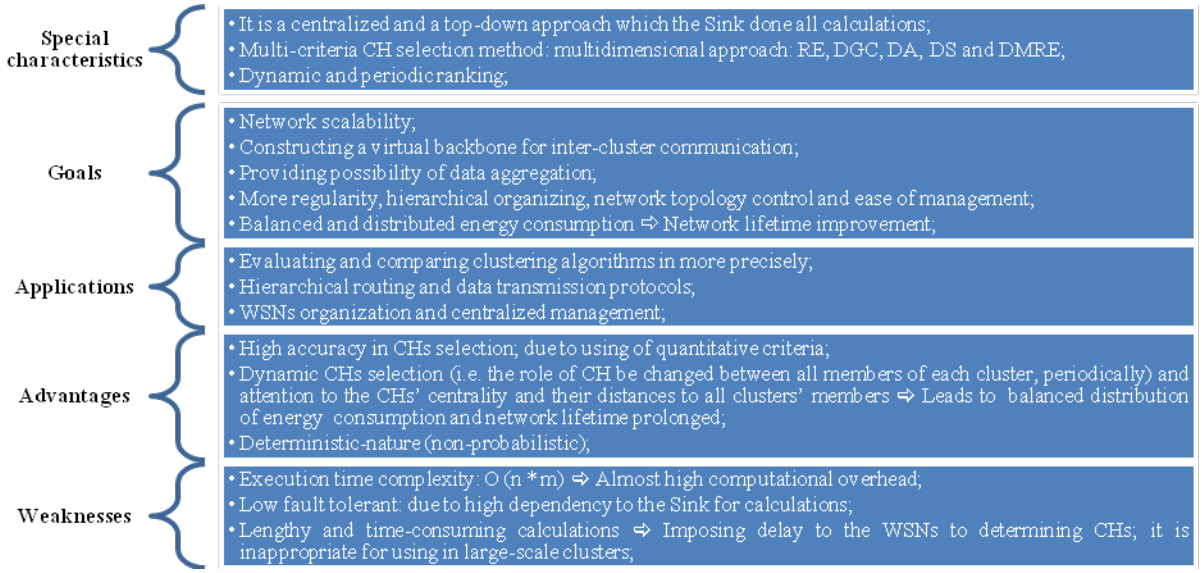


Fig. 7. Different aspects and properties of the proposed algorithm

It is better to using the proposed algorithm for small-scale WSNs with low clusters with low count of sensor nodes; or using it for large-scale WSNs with many clusters which each cluster has low count of sensor nodes.

Also, some general findings of this paper in WSNs' clustering area are as following:

- In large-scale WSNs, clustering leads to multi-hop communications ⇨ Multi-level cluster hierarchies ⇨ Preserving energy efficiency independent of the growth of the network.
- Further improvements in WSNs' reliability should consider possible modifications to the re-clustering mechanisms that following the initial CH selection. These modifications should be able to adapt the network clusters to maintain network connectivity while reducing the wasteful resources associated with periodic re-clustering.
- Clustering algorithms are different in many parameters such as CH selection criteria and clusters' formation methods.
- Clustering has three main phases, including of clusters formation, CHs selection and steady-state phase; so that: required time for (clusters formation and CHs selection phases) is much less than required time for (Steady-state phase).
- Some of most important challenges of clustering in WSNs are:
  - Clusters formation criteria and techniques;
  - CHs selection criteria and methods;
  - Overlapping and boundary nodes;
- Some problems of existent clustering algorithms which leading to their impracticality are as following:
  - Different assumptions, various clustering parameters, different operational environment and design constraints;
  - WSNs are large-scale networks;
  - Energy constraints; so, necessity of minimal message overhead;
- Clustering leads to hierarchical routing and data gathering, high scalability, data aggregation independent to the growth of the WSN, reducing communications, prolonging the WSN's lifetime and making more efficient usage of the critical resources.

There are several additional issues should be further studied in future researches. Some of most challenging proposed topics of these issues are including of:

- A technique for CHs balanced distribution in the operational environment (WSN);
- A method to calculating the optimal number and optimal size of clusters;
- A technique for estimation of the optimal rate of CHs rotation;
- Discussing on clusters overlapping and boundary nodes;
- Discussing on security attacks and threats in clustering based WSNs;
- A dynamic, energy-efficient and light-weight clustering (clusters formation and CH selection) criteria and algorithm for WSNs;

Finally, there is no universal clustering and CH selection algorithm which fits to different operational environments; because clustering algorithm is depending to the operational environments and its characteristics.

REFERENCES

[1] J. Yick, B. Mukherjee and D. Ghosal; Wireless Sensor Network Survey; Elsevier's Computer Networks Journal, 52 (2292-2330); 2008.

[2] H. Jadidoleslamy; A High-level Architecture for Intrusion Detection on Heterogeneous Wireless Sensor Networks: Hierarchical, Scalable and Dynamic Reconfigurable; International Journal of Wireless Sensor Network (WSN); Vol. 3. No. 7, pp. 241-261; 2011.

[3] S. Opricovic and G. H. Tzeng; Extended VIKOR method in comparison with outranking methods; European Journal of Operational Research, 178 (2), pp. 514-529; 2007.

[4] S. Opricovic and G. H. Tzeng; Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS; European Journal of Operational Research, 156 (2), pp. 445-455; 2004.

[5] K. K. Lai, S. Y. Wang and L. Yu; Progress in risk management guest editors introduction; International Journal of Information Technology & Decision Making, 5 (3), pp. 419-420; 2006.
[6] G. L. Fu, C. Yang and G. H. Tzeng; A multi-criteria analysis on the strategies to open Taiwan's mobile virtual network operators services; International Journal of Information Technology & Decision Making, 6 (1), pp. 85-112; 2007.
[7] A. A. Abbasi and M. Younis; A survey on clustering algorithms for wireless sensor networks; Computer Communications, Vol. 30, pp. 2826–2841; 2007.
[8] M. C. Thein and T. Thein; An Energy Efficient Cluster-Head Selection for Wireless Sensor Networks; International Conference on Intelligent Systems, pp. 287-291; 2010.
[9] A. Youssef, M. Younis, M. Youssef, and A. Agrawala; Distributed formation of overlapping multi-hop clusters in wireless sensor networks; in Proceedings of the 49th Annual IEEE Global Communication Conference; 2006.
[10] M. Ye, C. Li, G. Chen, and J. Wu; EECS: An energy efficient clustering scheme in wireless sensor networks; in Proceedings of IEEE International Performance Computing and Communications Conference (IPCCC'05), pp. 535–540, 2005.
[11] Y. Jin, L. Wang, Y. Kim, and X. Yang; EEMC: An energy-efficient multi-level clustering algorithm for large-scale wireless sensor networks; Computer Networks Journal, 52, pp. 542–562; 2008.
[12] G. Li and T. Znati; RECA: A ring-structured energy-efficient clustering architecture for robust communication in wireless sensor networks; International Journal Sensor Networks, 2(1/2), pp. 34–43; 2007.
[13] K. Yanagihara, J. Taketsugu, K. Fukui, S. Fukunaga, S. Hara, and K.I. Kitayama; EACLE: Energy-aware clustering scheme with transmission power control for sensor networks; Wireless Personal Communications, 40(3), pp. 401–415; 2007.
[14] P. Ding, J. Holliday, and A. Celik; Distributed energy efficient hierarchical clustering for wireless sensor networks; in Proceedings of the IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS05); 2005.
[15] J. Kamimura, N. Wakamiya, and M. Murata; A distributed clustering method for energy-efficient data gathering in sensor networks; International Journal on Wireless and Mobile Computing, 1(2), pp. 113–120; 2006.
[16] N.M. Abdul Latiff, C.C. Tsimenidis, and B.S. Sharif; Energy-aware clustering for wireless sensor networks using particle swarm optimization; in IEEE International Symposium PIMRC'07, pp. 1–5; 2007.
[17] J. Liaw, D. ChenYi and W. YiJie; The Steady Clustering Scheme for Heterogeneous Wireless Sensor Networks; Ubiquitous, Autonomic and Trusted Computing, UICATC '09. Symposia and Workshops on, pp. 336-341; 2009.
[18] L. Qing, Q. Zhu and M. Wang; Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks; Computer Communications, 29 (12), pp. 2230-2237; 2006.

AUTHOR BIOGRAPHY

H. Jadidoleslamy is a PhD candidate in Information Technology (IT)-Information Security at the Malekashtar University of Technology (MUT) in Tehran, Iran. He received his BSc degree in Information Technology (IT) engineering from the University of Sistan and Balouchestan (USB), Zahedan, Iran, in September 2009. He also has been received his MSc degree in Information Technology (IT) engineering from the University of Guilan, Rasht, Iran, in March 2011. His research interests are including of Computer Networks (especially Wireless Sensor Network), Information Security (by focusing on Intrusion Detection System) and E-Commerce. He may be reached at tanha.hossein@gmail.com or jadidoleslamy@gmail.com.

Categorization of ITIL® Tools

Kralik Lukas, Lukas Ludek

***Abstract*—** This paper responds to requirement to improve the orientation between offered SW, as ITIL® tools. There are really a lot of amount thus offered tools and very often leads to poor implementation of ITIL® on the basis of badly chosen tools. So this article aims to create dividing, which should facilitate choice of a suitable tool. Simultaneously, this division will serve for further work on creating a methodology for evaluation of ITIL® tools.

***Keywords*—** ITIL®, ITIL® tools, tools categorization, IT service support, ITIL implementation.

I. Introduction

WITH development of information and communication technologies (ICT) and their intrusion into all sectors, gaining management and delivery of IT services different dimension and meaning. The quality of providing or managing of IT services can greatly affect the operation or performance of the company. For this reason it was introduced, the now internationally acclaimed standard known as ITIL ®. It is an abbreviation for Information Technology Infrastructure Library. It is a set of concepts and practices that allow better planning, use and improve the use of IT, whether by the providers of IT services or by the customers.

The project originated in Great Britain in the mid 80s. Development of the first version lasted until 1995, and except of Great Britain it was applied and also used in the Netherlands. Since then undergone a series of changes so that it always match the current demands and conditions. Currently is ITIL ® in version 3 (ITIL ® V3) and consists of five key books (titles) - hence the name for the library:

1. Service Strategy
2. Service Design
3. Service Transition
4. Service Operation
5. Continual Service Improvement

According to the general definition of tool is a means of realizing certain activities, possibly used to communicate the results of that activity. The tool is tied to a specific technology or with some real technological or social procedure (or process). Based on this definition and the current version of ITIL® v3 can say that ITIL® is an arbitrary software tool which use leads to provably improve and streamline the providing and managing IT services. The only condition is that there must be a SW. It follows that as ITIL® tool can be used even standard office software. Everything stems from its use. Many SW is described as ITIL® tool, but if not properly used so labeling it as ITIL® tool is certainly not in place

The uses of ITIL® tools are complicated due to the wide range of offered tools and often very expensive. This caused and to a certain extent still causes small and medium companies are disinterest of the use of ITIL®. On the other hand, recently is beginning to discover significant amounts of Free and Open Source SW even between ITIL® tools.

II. SYSTEMIZATION OF ITIL® TOOLS

Due to the variety of software tools that support service management according to ITIL ® is very difficult to create and define a formal category for ITIL ® tools. The vast majority of software tools that are currently used in practice support a variety of processes. Tools focused on only one process is almost a matter of history. Categorization ITIL tools according to the current version of ITIL ® V3 is so complicated that the current version focuses on the management of IT services compared to the ITIL ® V2 was focused on the processes which allow easier categorization.

SW, which can be used as ITIL® tools, can be generally divided into three basic categories by (fig. 1):

1. Availability - way of licensing
2. Number of main functions
3. Main purpose

A. Division by availabilit

The simplest division ITIL® tools are according to availability or by the license under which it is available.

1. Proprietary SW
   - Commercial SW
   - Freeware
2. Open Source SW
3. Free SW

This work was supported by grant No. IGA/FAI/2014/020 from IGA (Internal Grant Agency) of Thomas Bata University in Zlin..

F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (corresponding author to provide phone: 303-555-5555; fax: 303-555-5555; e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar. colostate.edu).

1) Proprietary SW

Also known as a closed source software is software where its author modifies licenses (typically EULA) or otherwise, the possibility of its use. For such software is usually available free source code or it is impossible free to make modifications and distribute the resulting work.

a) Commercial SW

It is distributed for a fee. This means that if you want to use the product, you have to pay for creators. Such software usually can only be used by the limitation of its license. It is often limited by number of installations of software simultaneously, transfer, license or right to modification of the product.

b) Freeware

This type of software is distributed free of charge (or for a symbolic fee type of mission cards, often the author allows (but does not require) for the satisfaction of sending a donation), sometimes is talked about the type of software licenses. Conditions for the free use and redistribution are defined in the license agreement, which is often specific to each freeware.

The freeware author retains the copyright, for example, does not allow any program modification or restrict free use only for specific purposes (eg various combinations of the following restrictions: only for non-commercial purposes, only for personal use, only the home PC, only education in schools, only charities, only specific types of equipment, just to view files generated by the actual paid software, etc..). In some cases, the author also requires free registration or restricts the manner of distribution. Some freeware can also be used in companies working on computers, but only if it is not used for the direct providing of commercial services. Freeware software is so different from Free Software or Open Source software.

2) Free and Open Source SW

At first sight, the differences between Free and Open Source SW minimal and for layperson it is easy to swaps between these two types of SW. The main difference is the ideology of Gross.

a) Open Source SW

According to the Open Source Initiative [12], the SW must meet several requirements. These assumptions are not restricted as it could of Open Source associate only to the obligation to provide the purchaser access to the source code of a computer program, but also include other legal relations. These are the following requirements to be met by the license terms to a computer program (the definition of Open Source version 9.1):

- Free redistribution
- Source code
- Derived works
- Integrity of the author's source code
- No discrimination against persons or groups
- No discrimination against fields of endeavor
- Distribution of license
- License must not be specific to a product
- License must not restrict other software
- License must be technology-neutral

b) Free SW

Free software" means software that respects users' freedom and community. Roughly, it means that the users have the freedom to use this SW. Thus, "free software" is a matter of liberty, not price. To understand the concept, you should think of "free" as in "free speech," not as in "free beer".

A program is free software if the program's users have the four essential freedoms:

1. The freedom to run the program, for any purpose (freedom 0).
2. The freedom to study how the program works, and change it so it does your computing as you wish (freedom 1). Access to the source code is a precondition for this.
3. The freedom to redistribute copies so you can help your neighbor (freedom 2).

The freedom to distribute copies of your modified versions to others (freedom 3). By doing this you can give the whole community a chance to benefit from your changes. Access to the source code is a precondition for this.

B. Division by main purpose.

Based on the experience from practice ITIL tools can be divided into seven categories according to their primary purpose.

4. Service desk
5. Monitoring, event & remote management
6. Service life cycle
7. Service portfolio and management
8. Cloud
9. Information security
10. Others

1) Service desk

Service Desk is the single point of contact between the service provider and users. A typical Service Desk manages Incidents and service requests and handles communication not only with users but also with the management of the company. For its correct operation are needed different tools. They are mostly integrated into a single software solution most often in the form of a portal. However, there are a number of tools aimed at specific function or process (e.g. Service Level Management - SLM).

ISBN: 978-1-61804-236-1

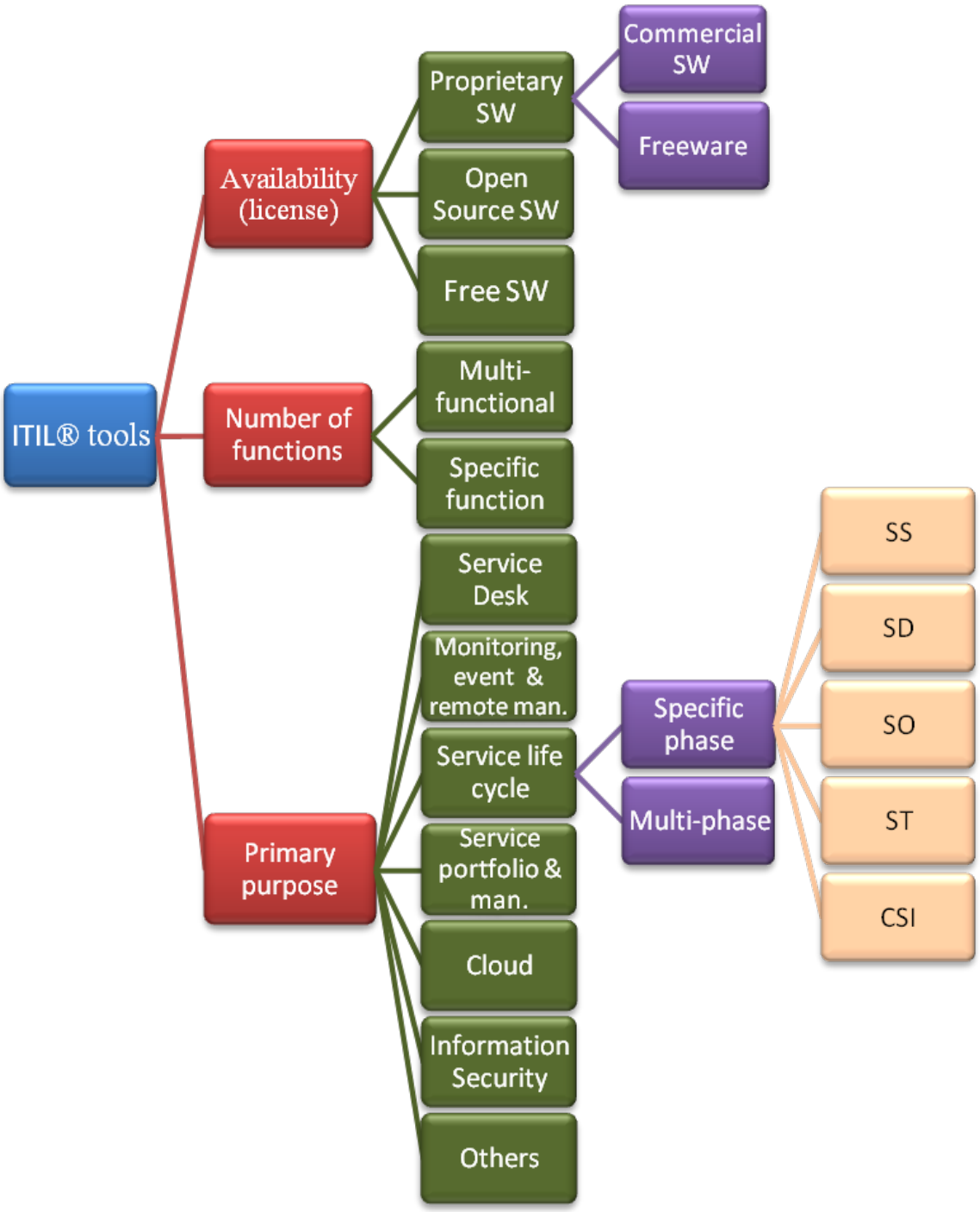


Fig. 1 Systemization and categorization of ITIL® tools

ISBN: 978-1-61804-236-1

2) Monitoring, event & remote management

Previously, these tools can be found under the name of NSM (Network and System Management). Allow monitoring networks and individual elements, systems, servers, applications and tracking incidents and other events by setting thresholds for optimal use of allocated resources and components. Although it is not a rule, it is integrated into the Incident Management and in most cases allows remote management.

3) Service life cycle

Specifically, it is a tool aimed at managing and supporting the entire lifecycle services. This area is also called the ALM (Application Lifecycle Management). But here come the tools of the field, which formally ITIL® does not cover (e.g. software development).

This type of instruments covering various platforms for developers, including support for versioning (source code revision tool), visualization platforms and different ways of testing (functional, security, load, ...) and both manual and automated.

4) Service portfolio and management

Tools in this category helps manage and control a complete portfolio of services, projects and programs. In addition, it is support a variety of processes such as Demand Management, Project Management, Program Management, Financial Management, Time Management and Resource Management.

5) Cloud

In this category are tools for the management and providing services in cloud for providers, as well as for users, or customers. Tools allow offer services (ordering), activation (deployment), their providing (provisioning) and of course invoicing (billing). However intervene here even instruments from category Service Desk and all functions are integrated into a single portal solution.

With taking into account to the events at present there is a great emphasis on speed, security, automation and intuitiveness of a particular solution. This category of instruments is typically proprietary software because they are designed for producers and their HW. However, Open Source software today has actually covers a wide range of areas and also for this category is not a problem to find a representative between Open Source and Free SW.

6) Information security

This category includes instruments starting with the anti-virus protection, through various tools for data security and test programs (penetration tests) to tools for monitoring. When monitoring is, however, an emphasis on security attributes (data theft, hacking, data corruption, etc.). Included in this category are access control systems (Access Management), which include central authentication and authorization of users, including the use directory services to control access to network elements, mobile devices. Finally, there are also physical security management, data protection and compliance with safety standards.

III. Conclusion

Update ITIL v3 has brought a number of changes. One of them is the approach to ITIL tools. This change led to simplify the implementation of ITIL, or allow IT managers to choose from a much wider range of software tools that can be considered as ITIL tools. Use of ITIL® tools is complicated and often very expensive due to the offered a wide range of tools. This caused and to a certain extent still causes small and medium companies are disinterest of the use of ITIL®. On the other hand, recently is beginning to discover significant amounts of Free and Open Source SW even between ITIL® tools.

The aim of dividing and systemization of ITIL® tools is make the orientation in the offered SW tools not only easier, but also to prove to IT managers working in small and medium-sized companies that the use of ITIL® tools and thus the implementation of ITIL® is not a matter for only large companies and international enterprises.

References

[1] Automated Unattended Installation in Kovárna Viva, a.s. In: International journal of computers. Oregon (USA): North Atlantic University Union, 2014, s. 7. ISSN 1998-4308.

[2] KRÁLÍK, Lukáš. Searching sources and evaluation criteria for open source itil® tools. In: Mezinárodní Masarykova Konference Pro Doktorandy A Mladé Vědecké Pracovníky. Hradec Králové: Magnimitas, 2013, s. 6. ISBN 978-80-87952-00-9.

[3] KRALIK, Lukas. Analysis for Automated Unattended Installation. In: Recent Advances in Automatic Control, Information and comunications: Proceedings of the 14th International Conference on Automation & Information (ICAI '13). Valencia (Španělsko): WSEAS press, 2013, s. 5. ISBN 978-960-474-316-2ISSN 1790-5117.

[4] KUFNER, Vladimír. ITIL V3: Změny v klíčových publikacích. DSM - data security management. 2012, č. 2, s. 7.

[5] BUCKSTEEG, Martin. ITIL 2011. 1. vyd. Brno: Computer Press, 2012, 216 s. ISBN 978-80-251-3732-1.

[6] ITIL continual service improvement [online]. 2nd ed. London: TSO, 2011, xi, 246 s. [cit. 2013-07-22]. Best Management Practice. ISBN 978-0-11-331308-2. Dostupné z: http://www.best-management-practice.com

[7] ITIL service transition [online]. 2nd ed. London: TSO, 2011, xii, 347 s. [cit. 2013-07-22]. Best Management Practice. ISBN 978-0-11-331306-8. Dostupné z: http://www.best-management-practice.com

[8] ITIL service design [online]. 2nd ed. London: TSO, 2011, xi, 442 s. [cit. 2013-07-22]. Best Management Practice. ISBN 978-0-11-331305-1. Dostupné z: http://www.best-management-practice.com

[9] ITIL service operation [online]. 2nd ed. London: TSO, 2011, xi, 370 s. [cit. 2013-07-22]. Best Management Practice. ISBN 978-0-11-331307-5. Dostupné z: http://www.best-management-practice.com

[10] ITIL: service strategy [online]. London: Stationery Office, 2011, xii, 264 s. [cit. 2013-07-22]. ISBN 978-011-3310-456. Dostupné z: http://www.best-management-practice.com/

Analogy of using intelligence and smart filters such as two stage Kalman in cloud computing

Mehdi Darbandi

Department of Electrical Engineering and Computer Science at Iran University of Science and Technology
mahdidarbandi@hotmail.com

***Abstract*—In this paper, at first we consider significant influences of this technology on some of the biggest companies and organizations all over the world, after that we present performance comparison of two stage Kalman filtering technique for surveillance permeating tracking in social networks such as cloud computing, we demonstrate mathematically all the equations and formula of this filter and also by the means of some MATLAB simulations we can use this technique for avoiding the entrance and existence of hacker and crackers [17].**

Keywords- Cloud Computing and Its Influences, Security, Kalman Filter, Evolutionary Algorithms, two stage Kalman estimator.

I. INTRODUCTION

Another important factor of such network is, suppose you are an inventor and you make new invention or wants to work on a special project. With use of cloud computing, you can share your project or your task with who you want. You can ask them about their suggestions and even complete your project with accompanying of them. Even you can define their level of access to different aspects of your own project. This feature is very important one for whom they release several new applications or things every day, they can share their application and all of the users can test it or even evaluate it and by the means of this feature, before the final releasing of your network you can find your application weaknesses.

In another word, cloud computing is doing processing and storage works on an online platforms, instead of processing on information and data inside of your computer you do your processes and duties on the clouds. With this groundbreaking capability, you're able to do your works more easily and cheaper. Because, everyday you do not need to concern about updating your software and hardware equipments.

Most of cloud services are providing through web browsers, so if we design and program secure and reliable web browsers we're able to intensify the security of our network. Also, we can implement new and intelligent applications on web browsers, by doing such actions; we're able to intensify security of our network.

Cloud computing is the best choice for the users with increasing demands. Suppose you are astronomer or even DNA scientists and you want to process on data's that attain from one source. With developing of science and technology, you need to develop your system capabilities all days, for example everyday new versions of software's come and/or you acquire images with higher resolutions and more precise. According to these extensive, you need to develop your system and specially your hardware, if you are the user of traditional users and you wants to update your system day by day, you should afford large amount of monies, but suppose if you are cloud user, every second you have the access to the latest versions of hardware and software's, because you use your necessities through internet platforms and connections.

Even on cloud platforms you can buy new application from one provider and you develop it and sell it to other users, through this work you can do business on cloud platforms and earning lots of money.

We don't have such problems in Cloud Computing, because every time you log out from your account and logging to it next time, you'll see the latest update of the software, without need to developing the hardware of your system, because the required hardware is provided by cloud resources. Also you don't pay additional costs for this software developing – buying licenses or pay for updating the software.

II. EVALUATING THE ROLE OF CLOUD COMPUTING ON THE FUTURE PRODUCTS OF INFORMATION TECHNOLOY INDUSTRY

Cloud computing has been around for some time – consider, for example, web-based personal email accounts – but it has primarily been used by

consumers. Today, a combination of sophisticated software, pervasive and interconnected devices, and ever-faster broadband connections is allowing governments and businesses to move beyond in-house IT systems to a more flexible model based on applications and services delivered over the Internet. These public and private sector users enjoy a range of benefits, including:

- Cost savings: Because the cloud frees users of the need to maintain their own IT infrastructure, they are able to spend their IT budgets more effectively and devote more of their human capital resources to their core business functions. Savings will increase as clouds grow: economists estimate that the combined impact of consolidating overhead and power costs and pooling computing resources can result in long-term savings of up to 80% when comparing large and small clouds.
- New opportunities for all: With cloud computing, organizations of any size and in virtually any location can tap into supercomputing power and software applications that previously were available only to the largest global companies. People also can build entirely new computing tools in the cloud.
- Increased agility and speed: Unprecedented computing power and storage capacity now available in the cloud allows organizations to roll out new applications and services with significantly greater speed – and less risk – than in the past. Services that once would have required large capital investments and lengthy deployments can be launched in a matter of weeks or even days.
- Reduced carbon footprint: Studies show that the cloud can produce real energy-efficiencies and reduce the carbon footprint of many business applications, thereby helping governments and industry achieve their green goals, reduce the environmental impact of IT, and enable a greener society [12].

In light of these benefits, it is not surprising that users are enthusiastic about the cloud. For example, KPMG in the Netherlands found last year that an overwhelming 59 percent of Dutch decision-makers and business leaders agree that "cloud computing is the future model of IT." Enthusiasm for the cloud is high in the U.S. as well. A survey conducted by Penn Schoen Berland for Microsoft found that 58 percent of consumers and 86 percent of senior business leaders in the U.S. are excited about the potential of cloud computing to change the way they use technology. The majority of consumers and business leaders believe these technologies can help government operate more efficiently and effectively as well [12].

That's the good news. The less good news is that almost every survey on the cloud also reveals that users are concerned about privacy and security. For example, a 2010 survey by the World Economic Forum found that 90 percent of respondents in Europe see privacy as a "very serious" constraint on adopting cloud computing. As people and organizations around the world move information from desktops to their mobile devices and into the cloud, they want to know that their data will remain safe and protected [12].

In Europe, Digital Agenda Commissioner Neelie Kroes has taken up this issue and urged the adoption of "clear and cloud-friendly rules. Because a 'cloud' without clear and strong data protection is not the sort of cloud we need." Likewise, the U.S. Department of Commerce recently observed that the ability to "safely use services such as cloud-based email and file storage to their full potential depends on privacy protections that are consistent with other computing models." We agree. Put simply, it is in the collective interest of all stakeholders that cloud users have well-founded confidence in the cloud.

Addressing privacy and security concerns in the cloud is industry's responsibility in the first instance. Microsoft fully embraces this responsibility, and the next section describes some of the many ways in which we are engaging with our customers to help them understand their rights and make informed choices when using cloud computing. Governments also have a critical role to play and the remaining sections of the paper propose steps that they can take to promote privacy and security in cloud computing, including updating legal frameworks to make clear whose laws apply – and how they apply – to data in the cloud and avoiding overly restrictive laws on the movement of data across borders. A proactive but balanced approach will best help spur innovation and drive resulting investment, job opportunities, and other benefits [12].

III. PERFORMANCE COMPARISON OF *TWO STAGE KALMAN FILTERING TECHNIQUE FOR SURVEILLANCE PERMEATING TRACKING IN CLOUD COMPUTING [17]:*

1. Statement of the Problem:

The problem of interest is described by the discretized equation set [15]:

$$X_{k+1} = A_k X_k + B_k U_k + W_k^x \quad (1)$$

$$U_{k+1} = C_k U_k + W_k^u \quad (2)$$

$$Z_k = H_k X_k + V_k \quad (3)$$

Where $X_k \in R^n$ is the system state, $U_k \in R^m$ and $Z_k \in R^p$ are the input and the measurement vectors, respectively. Matrices A_k, B_k, C_k and H_k are assumed to be known functions of the time interval k and are of appropriate dimensions. Matrix C_k is assumed nonsingular. The process noises W_k^x, W_k^u and the measurement noise V_k are zero-mean white Gaussian sequences with the following covariance's: $E[W_k^x (W_l^x)'] = Q_k^x \delta_{kl}$, $E[W_k^x (W_l^u)'] = Q_k^{xu} \delta_{kl}$, $E[W_k^u (W_l^u)'] = Q_k^u \delta_{kl}$ $E[V_k V_l'] = R_k \delta_{kl}$, $E[W_k^x V_l'] = 0$ and $E[W_k^u V_l'] = 0$, where $'$ denotes transpose and δ_{kl} denotes the Kronecker delta function. The initial states X_0 and U_0 are assumed to be uncorrelated with the sequences W_k^x, W_k^u and V_k. The initial conditions are assumed to be Gaussian random variables with $E[X_0] = \hat{X}_0$, $E[X_0 X_0'] = P_0^x$, $E[U_0] = \hat{U}_0$, $E[U_0 U_0'] = P_0^u$, $E[X_0 U_0'] = P_0^{xu}$.

Treating X_k and U_k as the augmented system state, the AUSKE is described by [15]:

$$X_{k+1|k+1}^{Aug} = X_{k+1|k}^{Aug} + K_{k+1}^{Aug}(Z_{k+1} - H_{k+1}^{Aug} X_{k+1|k}^{Aug}) \quad (4)$$

$$X_{k+1|k}^{Aug} = A_k^{Aug} X_{k|k}^{Aug} \quad (5)$$

$$K_{k+1}^{Aug} = P_{k+1|k}(H_{k+1}^{Aug})'[H_{k+1}^{Aug} P_{k+1|k}(H_{k+1}^{Aug})' + R_k]^{-1} \quad (6)$$

$$P_{k+1|k} = A_k^{Aug} P_{k|k}(A_k^{Aug})' + Q_k \quad (7)$$

$$P_{k+1|k+1} = (I - K_{k+1}^{Aug} H_{k+1}^{Aug}) P_{k+1|k} \quad (8)$$

Where

$$X_k^{Aug} = \begin{bmatrix} X_k \\ U_k \end{bmatrix}, \quad K_k^{Aug} = \begin{bmatrix} K_k^x \\ K_k^u \end{bmatrix}, \quad P_k = \begin{bmatrix} P_k^x & P_k^{xu} \\ (P_k^{xu})' & P_k^u \end{bmatrix},$$

$$A_k^{Aug} = \begin{bmatrix} A_k & B_k \\ 0_{m\times n} & C_k \end{bmatrix}, \quad H_k^{Aug} = \begin{bmatrix} H_k \\ 0_{p\times m} \end{bmatrix}', \quad Q_k = \begin{bmatrix} Q_k^x & Q_k^{xu} \\ (Q_k^{xu})' & Q_k^u \end{bmatrix}$$

Where the superscript 'Aug' denotes the augmented system state, I denotes the identity matrix of any dimension and $0_{m\times n}$ is a $m \times n$ zero matrix. It is clear from (4)-(8) that the computational cost of the AUSKE increases with the augmented state dimension. The OPSKE formulation is based on the following equations [15]:

$$\hat{\bar{X}}_{k+1|k+1} = \hat{\bar{X}}_{k+1|k} + K_{k+1}(Z_{k+1} - H_{k+1}\hat{\bar{X}}_{k+1|k}) \quad (9)$$

$$\hat{\bar{X}}_{k+1|k} = A_k \hat{\bar{X}}_{k|k} \quad (10)$$

$$K_{k+1} = P_{k+1|k}^x H_{k+1}'[H_{k+1} P_{k+1|k}^x (H_{k+1})' + R_k]^{-1} \quad (11)$$

$$P_{k+1|k}^x = A_k P_{k|k}^x (A_k)' + Q_k^x \quad (12)$$

$$P_{k+1|k+1}^x = (I - K_{k+1} H_{k+1}) P_{k+1|k}^x \quad (13)$$

$$N_{k+1} = [I - K_{k+1} H_{k+1}] M_{k+1} \quad (14)$$

$$\hat{U}_{k+1|k+1} = \hat{U}_{k+1|k} + K_{k+1}^u [\tilde{\bar{Z}}_{k+1} - H_{k+1} M_{k+1} \hat{U}_{k+1|k}] \quad (15)$$

$$\hat{U}_{k+1/k} = C_k \hat{U}_{k/k} \quad (16)$$

$$K_{k+1}^u = 2P_{k+1|k}^u M_{k+1}' H_{k+1}' \times [3H_{k+1} M_{k+1} P_{k+1|k}^u M_{k+1}' H_{k+1}' + P_{k+1|k}^z]^{-1} \quad (17)$$

$$P_{k+1|k+1}^u = P_{k+1|k}^u + 3K_{k+1}^u H_{k+1} M_{k+1} P_{k+1|k}^u M_{k+1}' H_{k+1}' (K_{k+1}^u)' + K_{k+1}^u P_{k+1|k}^z (K_{k+1}^u)' - 2P_{k+1|k}^u M_{k+1}' H_{k+1}' (K_{k+1}^u)' - 2K_{k+1}^u H_{k+1} M_{k+1} P_{k+1|k}^u \quad (18)$$

$$P_{k+1|k}^u = C_k P_{k|k}^u C_k' + Q_k^u \quad (19)$$

$$P_{k+1|k}^z = H_{k+1} P_{k+1|k}^x H_{k+1}' + R_{k+1} \quad (20)$$

$$P_{k+1|k}^{zu} = H_{k+1} M_{k+1} P_{k+1|k}^u \quad (21)$$

$$\hat{X}_{k+1|k} = \hat{\bar{X}}_{k+1|k} + M_{k+1} U_{k+1} \quad (22)$$

$$\hat{X}_{k+1|k+1} = \hat{\bar{X}}_{k+1|k+1} + N_{k+1} U_{k+1}$$

$$M_{k+1} = [A_k M_k + B_k] C_k^{-1}, \quad k = 2,3,.... \quad (23)$$

$$M_1 = B_0 C_0^{-1}$$

$$N_{k+1} = [I - K_{k+1} H_{k+1}] M_{k+1} \quad (24)$$

2. Performance Evaluations:

To demonstrate the computational advantage of the OPSKE over the AUSKE, the number of arithmetic operations are considered, i.e., multiplications and summations. The arithmetic operations of a standard Kalman estimator with state dimension n and measurement dimension p, are listed in Table 1. It is clear from the equations (4)-(8) and Table 1, that the arithmetic operations required for the AUSKE which has state dimension $n+m$ and measurement dimension p, are $M(n+m, p)$ for multiplications and $S(n+m, p)$ for summations. Table 2 shows the arithmetic operations of the input estimation and the auxiliary matrices needed by the OPSKE which has state dimension n, measurement dimension p and input vector dimension m. Note that the number of the arithmetic operations of the AUSKE increases with the augmented state dimension, which makes the algorithm computationally inefficient. In contrast, the OPSKE based on the two-

stage decoupling technique required fewer computations. The efficiency of the OPSKE is due to order reduction, i.e., implementing two less order n and m partitioned filters. This enables the proposed algorithm to have much better computational efficiency than the AUSKE. So, the arithmetic operations required (AOR) for the AUSKE are [15]:

$$\begin{aligned} &AOR(AUSKE) = M(n+m,p) + S(n+m,p) \\ &= [3(n+m)^3 + 2(n+m)^2 p + 2(n+m)p^2 + p^3 + (n+m)^2 + 2(n+m)p] \\ &+ [3(n+m)^3 + 2(n+m)^2 p + 2(n+m)p^2 + p^3 - (n+m)^2 - (n+m)] \end{aligned} \quad (25)$$

The arithmetic operations required for the input estimation and auxiliary matrices, by the OPSKE as shown in Table 2 and using equations (15)-(24) are

$$\begin{aligned} &AOR(OPSKE) \\ &= M(n,p) + S(n,p) + M^{OP}(n,m,p) + S^{OP}(n,m,p) \\ &= [3n^3 + 2n^2 p + 2np^2 + p^3 + n^2 + 2np] \\ &+ [3n^3 + 2n^2 p + 2np^2 + p^3 - n^2 - n] \\ &+ [3mp + 2m^2 + 2m^2 p + 2mp^2 + p^3 + p^2 \\ &+ 4m^3 + 2n^2 p + 2nm + n^2 m + nm^2 + nmp] \\ &+ [-mp - m^2 - m + 2m^2 p + 2mp^2 + p^3 + 4m^3 \\ &+ 2n^2 p - 2np + p^2 - n + 2n^2 m + nm^2 + nmp] \end{aligned} \quad (26)$$

Using (25) and (26), the operational savings, denoted by OS_{AUSKE}^{OPSKE}, of the OPSKE as compared to the AUSKE are [17]:

$$\begin{aligned} &OS_{AUSKE}^{OPSKE} = AOR(AUSKE) - AOR(OPSKE) = \\ &M(n+m,p) + S(n+m,p) - M(n,p) \\ &- S(n,p) - M^{OP}(n,m,p) - S^{OP}(n,m,p) \\ &= -2m^3 + 15n^2 m + 17nm^2 - 4n^2 p + 6nmp \\ &- 2p^3 + 2np + n - m^2 - 2p^2 - 2nm \end{aligned} \quad (27)$$

And the operational savings of the OTSKE over the AUSKE are:

$$\begin{aligned} &OS_{AUSKE}^{OTSKE} = AOR(AUSKE) - AOR(OTSKE) = -4m^3 + \\ &12n^2 m + 12nm^2 + 4nmp + m - 2m^2 - p^3 - 2nm \end{aligned} \quad (28)$$

Therefore, using (27) and (28) the operational savings of the OPSKE over the OTSKE are:

$$\begin{aligned} &OS_{OTSKE}^{OPSKE} = AOR(OTSKE) - AOR(OPSKE) = 2m^3 + 3n^2 m \\ &+ 5nm^2 - 4n^2 p + 2nmp - p^3 + 2np + n - m + m^2 - 2p^2 \end{aligned} \quad (29)$$

It is clear from (27) and (29) that for m and $p \le n$, the proposed scheme has computational advantage over the AUSKE and it is comparable to the OTSKE. The operational savings discussed here will be tested as an example in the simulation results section. To measure the relative operational savings of the OPSKE with respect to the arithmetic operation required by the AUSKE ($AOR(AUSKE)$), the percentage of the operational savings defined as below:

$$POS_{AUSKE}^{OPSKE} = \frac{OS_{AUSKE}^{OPSKE}}{AOR(AUSKE)} \times 100 \quad (30)$$

Using (27), (29) and (30), the operational savings and the percentage of the operational savings, of the OPSKE comparing to the OTSKE and the AUSKE for different values of n, m and p are shown in Table 3. It can be inferred from Table 3 that the OPSKE has better overall performance than the AUSKE (averaged 32%) and the OTSKE (averaged 7.3%) [15].

Table 1:Standard Kalman Estimator Arithmetic Operation Requirements

	Variable	Number of Multiplications, $M(n,p)$	Number of summations, $S(n,p)$
1	$X_{k+1\|k+1}$	$2np$	$2np$
2	$X_{K+1\|k}$	n^2	$n^2 - n$
3	K_{k+1}^x	$n^2 p + 2np^2 + p^3$	$n^2 p + 2np^2 + p^3 - 2np$
4	$P_{K+1\|k}^x$	$2n^3$	$2n^3 - n^2$
5	$P_{K+1\|k+1}^x$	$n^3 + n^2 p$	$n^3 + n^2 p - n^2$
	Totals	$3n^3 + 2n^2 p + 2np^2 + p^3 + n^2 + 2np$	$3n^3 + 2n^2 p + 2np^2 + p^3 - n^2 - n$

Table 2:Input Estimation and Auxiliary Matrices Arithmetic Operation Requirements for the OPSKE

	Variable	Number of Multiplications $M^{OP}(n,m,p)$	Number of summations $S^{OP}(n,m,p)$
1	$U_{k+1\|k+1}$	$2mp$	$2mp$
2	$U_{K+1\|k}$	m^2	$m^2 - m$
3	K_{k+1}^u	$m^2 p + 2mp^2 + p^3 + p^2 + mp$	$m^2 p + 2mp^2 + p^3 - 2mp$
4	$P_{K+1\|k}^u$	$2m^3$	$2m^3 - m^2$

5	$P^{u}_{K+1\|k+1}$	$m^3 + m^2 p + m^2$	$m^3 + m^2 p - m^2$
6	$P^{z}_{k+1\|k}$	$2n^2 p$	$2n^2 p - 2np + p^2$
7	$\hat{X}_{k+1\|k}$	mn	mn
8	$\hat{X}_{k+1\|k+1}$	mn	$mn - n$
9	M_{k+1}	$n^2 m + m^3 + nm^2$	$n^2 m + m^3 + nm^2 - nm$
10	N_{k+1}	$n^2 m$	$n^2 m - nm$
11	$H_{k+1} M_{k+1}$	nmp	$nmp - mp$
	Totals	$3mp + 2m^2 + 2m^2 p + 2mp^2 + p^3 + p^2$ $+ 4m^3 + 2n^2 p + 2nm + n^2 m + nm^2 + nmp$	$-mp - m^2 - m + 2m^2 p + 2mp^2 + p^3 + 4m^3$ $+ 2n^2 p - 2np + p^2 - n + 2n^2 m + nm^2 + nmp$

Table 3:the Operational Savings and the Percentage of the Operational Savings of the OPSKE Compared to the AUSKE and the OTSKE

The state vector dimensions	OS^{OPSKE}_{AUSKE}	POS^{OPSKE}_{AUSKE} (%)	OS^{OPSKE}_{OTSKE}	POS^{OPSKE}_{OTSKE} (%)
$n=4, m=4, p=2$	1340	35.7	592	15.7
$n=4, m=2, p=2$	578	33.7	102	5.9
$n=4, m=2, p=1$	553	37.5	155	10.5
$n=4, m=1, p=1$	242	27.5	23	2.6
$n=4, m=3, p=3$	978	32.7	247	8.2
$n=10, m=2, p=2$	2954	25.1	132	1.12
Average	$\cong 1107$	32.0	$\cong 208$	7.3

3. Simulation Results:

To evaluate the proposed algorithm, an example of maneuvering target tracking problem which turns, in two-dimensional space is simulated such as permeating a hacker into a very important network or databases. In this simulation example, the performance of the OPSKE for the maneuvering target tracking has been compared with the traditional works that done in this concept, as an example of the AUSKE method. As mentioned before in the augmented state method the state vector includes the input vector i.e., acceleration and jerk parameter in maneuvering target tracking problem. The sampling interval is T=0.01 (sec) and target maneuver is applied at 9th second (900th sample). The initial conditions are selected similar for the AUSKE as well as the OPSKE. The state vectors are

$X_k = \begin{bmatrix} x_k & v_k^x & y_k & v_k^y \end{bmatrix}'$, $U_k = \begin{bmatrix} u_k^x & j_k^x & u_k^y & j_k^y \end{bmatrix}'$,

$X_k^{Aug} = \begin{bmatrix} x_k & v_k^x & y_k & v_k^y & u_k^x & j_k^x & u_k^y & j_k^y \end{bmatrix}'$

Where x_k, v_k^x, u_k^x and j_k^x denote the position, velocity, acceleration and jerk of the target along the x axis, respectively. We consider the target initial conditions for the state and the acceleration vectors as below [15]:

$X_0 = \begin{bmatrix} 2165\,m & -80\,m/s & 1250\,m & 25\,m/s \end{bmatrix}'$,

$U_0 = \begin{bmatrix} 0\,g & 0\,g/sec & 0\,g & 0\,g/sec \end{bmatrix}'$

$X_0^{Aug} = \begin{bmatrix} 2165\,m & -80\,m/s & 1250\,m & 25\,m/s & 0\,g & 0\,g/\sec & 0\,g & 0\,g/\sec \end{bmatrix}'$

The target begins to maneuver as $U_{900} = \begin{bmatrix} 0g & -0.7\,g/\sec & 0g & 0.4\,g/\sec \end{bmatrix}'$ for $9\ (\sec) \le t \le 90\ (\sec)$.

The system matrices are given by

$$A_k = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad B_k = \begin{bmatrix} T^2/2 & T^3/6 & 0 & 0 \\ T & T^2/2 & 0 & 0 \\ 0 & 0 & T^2/2 & T^3/6 \\ 0 & 0 & T & T^2/2 \end{bmatrix},$$

$$C_k = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad H_k = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}'$$

$$Q_k^u = 2\alpha\sigma_j \begin{bmatrix} T^3/3 & T^2/2 & 0 & 0 \\ T^2/2 & T & 0 & 0 \\ 0 & 0 & T^3/3 & T^2/2 \\ 0 & 0 & T^2/2 & T \end{bmatrix},$$

$$Q_k^x = 2\alpha\sigma_j \begin{bmatrix} T^7/252 & T^6/72 & 0 & 0 \\ T^6/72 & T^5/20 & 0 & 0 \\ 0 & 0 & T^7/252 & T^6/72 \\ 0 & 0 & T^6/72 & T^5/20 \end{bmatrix}$$

$$Q_k^{xu} = 2\alpha\sigma_j \begin{bmatrix} T^5/30 & T^4/24 & 0 & 0 \\ T^4/8 & T^3/6 & 0 & 0 \\ 0 & 0 & T^5/30 & T^4/24 \\ 0 & 0 & T^4/8 & T^3/6 \end{bmatrix}, \quad P_0^x = 10I_{4\times4},$$

$$P_o^u = 0.1I_{4\times4}, \quad P_0^{xu} = I_{4\times4}, \quad H_k^{Aug} = \begin{bmatrix} H_k \\ 0_{2\times4} \end{bmatrix}'$$

$$A_k^{Aug} = \begin{bmatrix} A_k & B_k \\ 0_{4\times4} & C_k \end{bmatrix}, \quad Q_k = \begin{bmatrix} Q_k^x & Q_k^{xu} \\ (Q_k^{xu})' & Q_k^u \end{bmatrix}, \quad P_k = \begin{bmatrix} P_k^x & P_k^{xu} \\ (P_k^{xu})' & P_k^u \end{bmatrix}.$$

Where $\sigma_j = 0.09(ms^{-3})$ the variance of the target is jerk and $\alpha = 0.0123\,(\mathrm{s}^{-1})$ is the reciprocal of the jerk time constant $\tau = 1/\alpha$. The measurement standard deviations of x and y target positions are: $\sigma_x = 10\sqrt{10}\,(m)$, $\sigma_y = 20\,(m)$. Thus, the measurement covariance matrix is $R_k = \begin{bmatrix} 1000 & 0 \\ 0 & 400 \end{bmatrix}$ for both methods [15]. The Root Mean Square Error (RMSE) index is used for the results evaluation.

Fig. 1 shows the actual value and the estimation of x and y and RMS errors of x and y positions estimations by the proposed OPSKE and the AUSKE. Fig. 2 shows the actual value and the estimations of v^x, v^y and the RMS errors of the x and y velocities estimations by the proposed method compared with the augmented method. The actual value and the accelerations estimations in the x and y directions and their corresponding averaged RMS errors can be seen in Fig. 3.Fig. 4 displays the actual value and the estimated jerk parameters are evaluated by the OPSKE and the AUSKE methodologies [15].

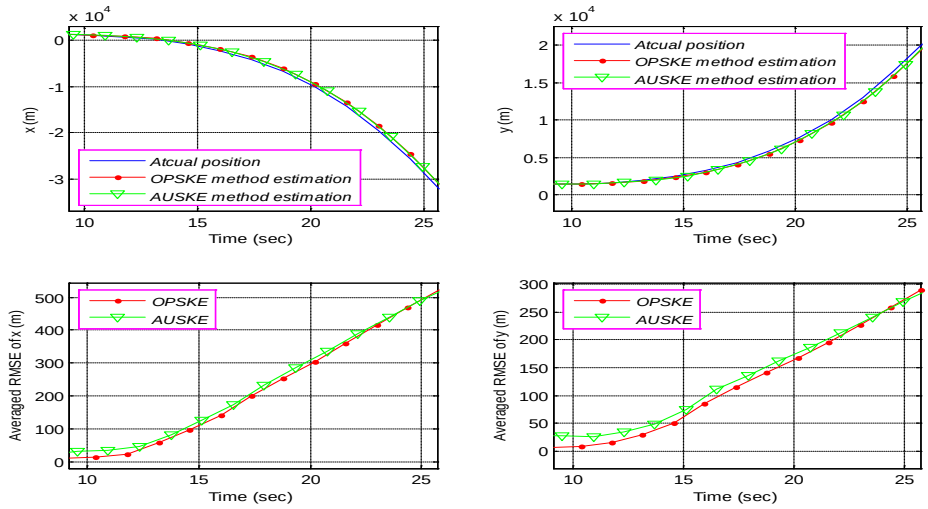


Fig. 1. The actual value and the estimation of the x, y positions and RMS errors estimations by the OPSKE and the AUSKE methods.

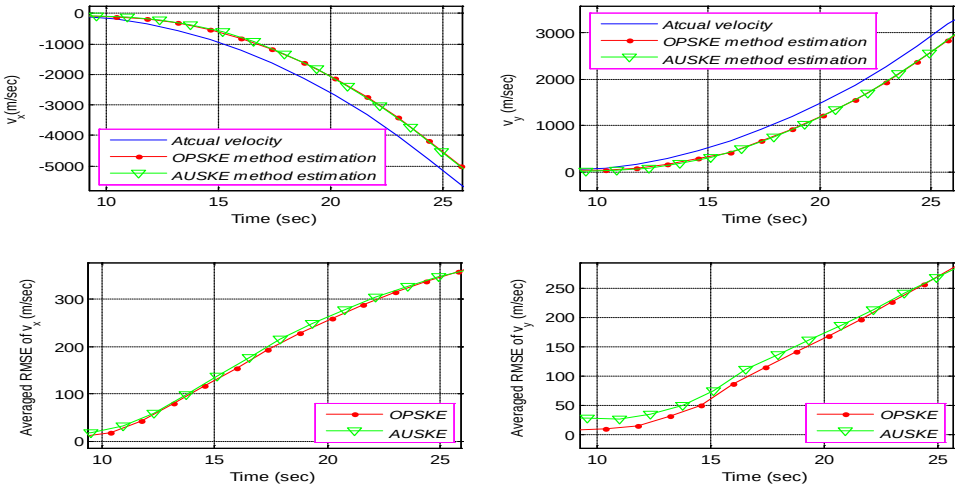


Fig. 2. The actual value and the estimation of v^x, v^y and RMS errors of x and y velocities estimations by the OPSKE and the AUSKE methods.

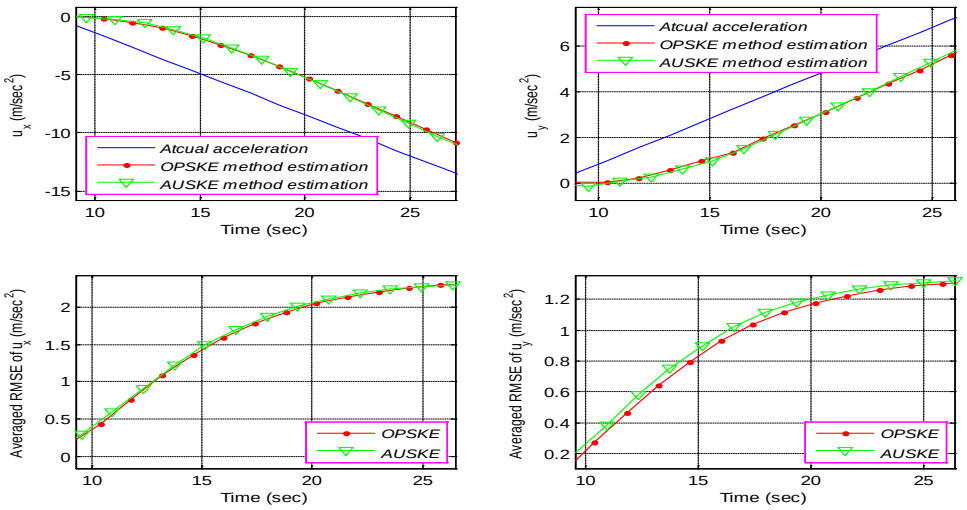


Fig. 3. The actual value and the estimation of acceleration in x and y directions and corresponding RMS errors by the proposed method compared with the augmented methods.

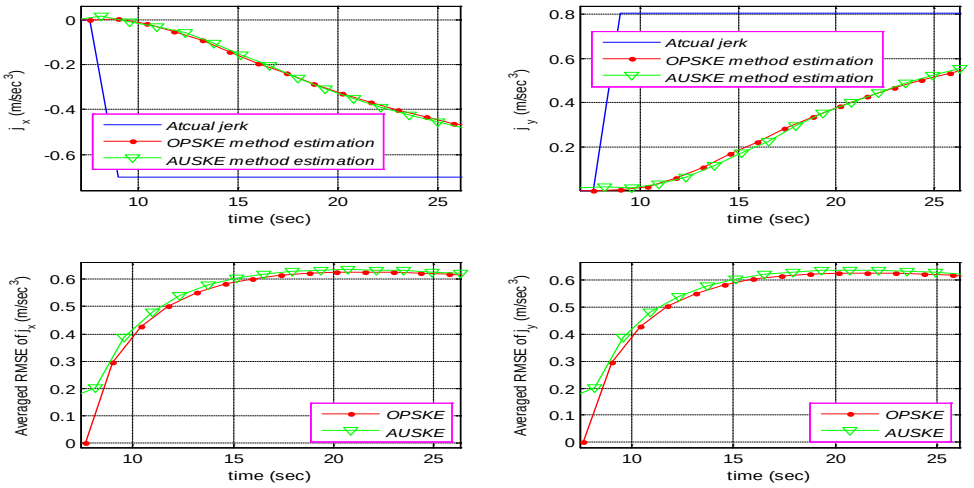


Fig. 4. The actual value and the estimation of jerk parameters and RMS errors by the OPSKE method compared with the AUSKE method.

ISBN: 978-1-61804-236-1

It is clear that the performance of the proposed OPSKE is as well as the results obtained by the AUSKE in the maneuvering target tracking problem. Note that in this example $n=4$, $m=4$ and $p=2$, and the operation savings for the OPSKE over the AUSKE and the OTSKE as shown in Table 3 are 1340 (or 35.7%) and 592 (or 15.7%), respectively.

IV. CONCLUSION

In this paper at first we review some basic definitions and principles of cloud computing and tell about different aspects of this technology. After that we talk about its influences on different industries and other technologies. Also, we talk about some of future works that will be done by huge and well-known companies, such as Microsoft, based on cloud platforms.

In the third section of our paper, we purpose new Kalman estimator which can be used for estimation and prediction purposes in cloud computing or even if we wants to track and trace hackers or crackers we can use such filter on such networks.

References

[1] Mehdi Darbandi "Applying Kalman Filtering in solving SSM estimation problem by the means of EM algorithm with considering a practical example"; published by the Journal of Computing – **Springer**, 2012; USA.

[2] Mehdi Darbandi; "Comparison between miscellaneous platforms that present for cloud computing and accreting the security of these platforms by new filter"; published by the Journal of Computing – **Springer,** 2012; USA.

[3] Mehdi Darbandi; "New and novel technique in designing electromagnetic filter for eliminating EMI radiations and optimization performances"; published by the Journal of Computing - **Springer**, 2012; USA.

[4] Mehdi Darbandi; "Appraising the role of cloud computing in daily life and presenting new solutions for stabilization of this technology"; published by the Journal of Computing - **Springer**, 2012; USA.

[5] Mehdi Darbandi; "Cloud Computing make a revolution in economy and Information Technology"; published by the Journal of Computing - **Springer**, 2012; USA.

[6] Mehdi Darbandi; "Considering the high impact of gettering of silicon on fabrication of wafer designing and optimize the designing with new innovative solutions"; published by the Journal of Computing – **Springer**, 2012; USA.

[7] Mehdi Darbandi; "Developing concept of electromagnetic filter design by considering new parameters and use of mathematical analysis"; published by the Journal of Computing - **Springer**, 2012; USA.

[8] Mehdi Darbandi; "Is the cloud computing real or hype Affirmation momentous traits of this technology by proffering maiden scenarios"; published by the Journal of Computing – **Springer**, 2012; USA.

[9] Mehdi Darbandi; "Measurement and collation overriding traits of computer networks and ascertainment consequential exclusivities of cloud computing by the means of Bucy filtering"; published by the Journal of Computing - **Springer**, 2012; USA.

[10] Mehdi Darbandi; "Unabridged collation about multifarious computing methods and outreaching cloud computing based on innovative procedure"; published by the Journal of Computing - **Springer**, 2012; USA.

[11] Mehdi Darbandi; "Scrutiny about all security standards in cloud computing and present new novel standard for security of such networks"; published by the Journal of Computing - **Springer**, 2012; USA.

[12] Microsoft's Accessible Technology Vision and Strategy; September 2011.

[13] MSc. Thesis of *Eman A. Aldakheel; College of Bowling Green; December 2011.*

[14] *www.wikipedia.org*

[15] A. Karsaz, H. Khaloozade, M. Darbandi; "Performance Comparison of the two-stage Kalman filtering Techniques for Target Tracking" Int. IEEE Conf. Harbin, China.

BIOGRAPHIES:

Mehdi Darbandi received his B.Sc. degree in Electrical Engineering from University of Mashhad in 2012. His research areas are Kalman Filter, Matlab Simulink, Evolutionary Algorithms, and Cloud Computing. He is now Master student at Iran University of Science and Technology (IUST); Tehran, Iran.

Knowledge Management Approaches for Business Intelligence in Healthcare

Nadia Baeshen

Abstract **-** Knowledge management (KM) is an intelligent process by which the gathered raw data is transformed into knowledge. KM have become an efficient approach for building practical and intelligent decision support systems in medical and healthcare domains. Business intelligence (BI) is a structured approach to preparing and using information to drive business activity and it is instrumental in turning raw data into knowledge that can be used to derive value .It can described as a value proposition that helps organization s in their decision making processes .This paper discusses two important knowledge management approaches that are used for business intelligence in the context of healthcare domain, namely; expert systems and data mining techniques.

Keywords **-** Business intelligence, Data mining, Expert systems, Healthcare, Knowledge management.

I. INTRODUCTION

During recent decades, knowledge is the aspiring elementary resource mandatorily required by all intelligent information processing systems. Knowledge engineers use artificial intelligence concepts and techniques to knowledge-based decision support systems. Knowledge management (KM) is emerging as the new discipline that provides the mechanisms for systematically managing the knowledge that evolves with the enterprise. Most large organizations have been experimenting with knowledge management with a view to improving profits, being competitively innovative, or simply to survive [1, 2]. Furthermore, exploiting technology enables organizations to derive knowledge from data and information collected as the business proceeds. It then may be exploited in decision making, product development, human resourcing, customer relationships, the supply chain and so on. Clearly, knowledge management needs to infiltrate every aspect of the enterprise to improve business efficiency. Most literature on KM classifies knowledge into two main categories: explicit knowledge and tacit knowledge. Explicit knowledge can be defined as things that are clearly stated or defined, while tacit knowledge can be defined as things that are not expressed openly, but implied [3, 4].

Business Intelligence (BI) is an umbrella term for various business managing approaches based on well-informed decisions, which lead to a high performance level within organizations. The Data Warehouse Institute defines BI as "a set of concepts and methodologies to improve decision making in business through the use of facts and fact-based systems"[5]. BI can be thought of as getting the right information to the right people at the right time and place to enable fact-based decisions. Recently two distinct understandings of the term BI (respectively BI system) exist – a data-centric and a process-centric. The data-centric position uses BI systems to combine operational data with analytical tools to present complex and competitive information to planners and decision makers. The objective is to improve the time liness and quality of inputs to the decision process [6]. BI is therefore mainly used to understand the capabilities available in the organization [7]. The process-centric position notes a major shortcoming in this inherent data-centricity. Because the collection, transformation, and integration of data as well as information supply and analysis are commonly isolated from business process execution, a great part of the information that intrinsically exits within an organization remains either unused or is at most partially used but deprived of its interpretation context [8]. As they see an organization as a set of well-integrated processes [9], BI therefore should be used to integrate the information world with the process world in order to facilitate decision making with an all-embracing information basis.

In the context of BI, knowledge management techniques can be seen as enabler for managing, storing, analyzing, visualizing, and giving access to a great amount of data. For this purpose, a wide range of intelligent technologies (e.g. expert systems, online analytical processing, data mining and knowledge discovery, grid computing, cloud computing) are used in developing of a BI systems. Technology is required to provide an integrated view of both, internal and external data (for example by means of a data warehouse). It is therefore the base for BI. The aim of this study was to discuss the benefits of the well known knowledge management approaches, namely; expert systems (ESs) and data mining (DM) from the business intelligence point of view in the context of healthcare domain.

II. BUSINESS INTELLIGENCE APPROACH IN HEALTHCARE ENVIRONMENT

From the BI perspective , we have the following three the main healthcare processes; (a) **Medical processes**, (b) **Business processes**, and (c) **Support processes.**

(a) Medical processes are those activities and work practices within a health care organization which are mainly

focused on the health services delivery ,e.g.; diagnostic and therapy , research and teaching , and nurse care.

(b) Business processes comprise activities that are needed to effectively run the health care organization and may not be, or only partially sector specific ,e.g; monitoring and controling , financial accounting , compliance and risk management , and Organizational Development .

(c) Support processes are used from both kinds of processes but only have an indirect impact on medical and business activities, e.g; communication ,human resources and logistics and supply

Based on our analysis for the recent publications during the last five years, one can conclude that, the BI serves an increasingly wide variety of departments in the provider market with an assortment of unique reporting and analysis applications. A robust BI environment offers healthcare organizations a host of business benefits including:

1. The ability to optimize resources (including physical space, equipment and devices, staff and supplies) in individual departments such as Surgical Services.
2. The ability to develop and monitor key performance indicators and clinical indicators to improve performance and quality.
3. The ability to conduct planning, budgeting, and forecasting more efficiently and accurately across large organizations.
4. The ability to effectively understand and manage the supply chain and logistics to contain costs and ensure consistent supply.
5. The ability to better ensure patient safety through efficient diagnostics and the identification and enforcement of best practice treatment protocols.
6. The ability to contain costs and improve performance and quality through human resources management and physician profiling

III. Expert Systems Approach in the Context of Business intelligence in Healthcare

Expert system (ES) is a consultation intelligent system that contains the knowledge and experience of one or more experts in a specific domain that anyone can tap as an aid in solving problems [10]. The most commonly systems are rule-based expert systems (RES) and case-based expert systems (CES). In RES the knowledge base stores the knowledge in the form of production rules (if-then statements). The inference engine contains a set of formal logic relationships which may or may not resemble the way that real human expert reach conclusions. CES uses case-based reasoning (CBR) methodology in which the system can reason from analogy from the past cases. This system contains what is called "case-memory" which contains the knowledge in the form of old cases (experiences). CES solves new problems by adapting solutions that were used for previous and similar problems [11]. The technology of CBR directly addresses the problems found in rule-based technology, namely: knowledge acquisition, performance, adaptive solution, maintenance.

In the last years various machine learning (ML) techniques have been proposed by the researchers in order to develop efficient biomedical knowledge-based systems[12,13].ML is an intelligent technique that tries to find a mathematical model that maps between inputs and outputs of a domain problem. There are two stages of using ML techniques. These are creating the mathematical model by learning mappings between given input and output. The second stage is using the model to predict an output, given unseen input.ML techniques offer a robust computational intelligence methods and algorithms that can help solving management problems in healthcare domains.

Based on our analysis of the recent publications during the last five years, one can summarized the benefits of the ESs technology to healthcare sector in the following;

a) Treatment choice – may be easier with the use of if-then rules of an expert system; Following the rules, a physician is able to infer treatment adequate to symptoms and/or to a specific illness;

b) Diagnosis support – this comes both from rule-based systems as well from case based ones. If-then rules enable encoding of knowledge linking symptoms to illnesses, while case-based reasoning enables finding the illness by comparing patients' symptoms to these stored in case-based knowledge base;

c) Analysis of treatment options – rule-based knowledge enables a so-called what-if analysis: what is probable to happen if we use a specific treatment?

d) Keeping medical history – is easy with case-based expert systems, where individual patients' cases may be stored both for statistical purposes and for case-based reasoning.

IV. Intelligent Data Mining Approach as a Business Intelligence Value Chain

Data mining approach is a complex process of using historical databases to improve subsequent decision making.DM methodology aims to extract useful knowledge and discover some hidden patterns from huge amount of databases which statistical approaches cannot discover [14]. Knowledge discovery in databases (KDD) process involves the following processes; (a) using the database along with any required selection, preprocessing, sub-sampling, and transformations of it, (b) applying data mining methods to enumerate patterns from it, and (c) evaluating the products of data mining to identify the subset of the enumerated patterns deemed knowledge. The data mining components of the KDD process is concerned with the algorithmic means by which patterns are extracted and enumerated from data (e.g. rough sets, fuzzy logic, neural networks).Data mining is supported by a host that captures the character of data in several different ways ,e.g. clustering, regression models, classification, summarization, link analysis and sequence analysis [15].

Fig. 1 shows the main functional phases of the medical knowledge discovery process[16,17].The preprocessing phase is often referred to as data cleaning. The cleaned data are stored in the warehouse. This is followed by data mining phase and its results are provided to an output generator (visualization) producing reports, action lists, or monitor reports. Each phase is supported by different methodologies. Data mining itself exhibits a plethora of algorithmic tools such as statistics, regression models, neural networks, fuzzy sets and evolutionary model. From fig. 1, it can be seen that, the knowledge discovery process is dynamic, highly interactive, iterative, and fully visualize able. Its main goals are to: (a) extract useful reports (b) spot interesting events and trends (c) support decision-making processes (d) exploit the data to achieve scientific, business, or operational goals.

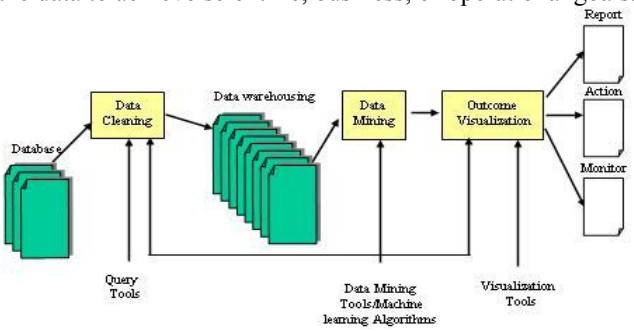


Fig 1: knowledge discovery process.

Data mining is supported by a host that captures the character of data in many different ways ,e.g. Classification, Clustering, Link analysis, Sequence Analysis, Regression Models, and Summarization. So, based on these intelligent techniques, data mining approach can performs and provides several tasks and benefits for the healthcare sector, e. g;

(a) Clustering and classification tasks are useful for statistical purposes,

(b) Link analysis – may support accurate diagnosis (by showing links between symptoms and illnesses) and efficient treatment – by revealing links between illnesses and medical drugs,

(c) Storing medical history in a warehouse may be useful for statistical analyses, and

(d) Discovering patterns and trends – used for diagnosis and reporting purposes.

V. DISCUSSION

Business Intelligence paradigm can be described as a value proposition that helps organizations in their decision-making processes [18]. Following to Porter [19], a value chain is a systematic approach to examine the development of competitive advantage, consisting of a series of activities that create and build value. All the stages and relationships in this approach will add value to the decision support process. Based on the introduced value chain, tasks like business analysis, enterprise reporting and performance management are possible.

Healthcare organization uses the business intelligence solution not only for analysis but also to change business processes and drive toward the value-driven healthcare vision. Business intelligence provides an integrated view of data that can be used to monitor key performance indicators, identify hidden patterns in diagnosis, illuminate anomalies in processes, and identify variations in cost factors, all of which facilitate accountability and visibility and can drive an organization towards efficiency.

From the technical point of view, healthcare-based business intelligence systems are complex to build, maintain and face the knowledge-acquisition difficulty. Efficiency of such systems is determined by the efficiency of the knowledge management techniques and methodologies. Knowledge management techniques provide an effective knowledge computing methods and robust environment for business intelligence in the healthcare decision making domain.

VI. CONCLUSION

This paper discusses two important knowledge management approaches that are used for business intelligence in the context of healthcare domain, namely; expert systems and data mining techniques. Both techniques can be seen as enabler for managing, storing, analyzing, visualizing, and giving access to a great amount of data .Intelligent data mining approach fits the " Value Chain" model of business intelligence "From DATA To PROFIT" approach .Biomedical data is transformed into relevant and useful information. Further, the obtained valuable knowledge supports any decision-making processes in order to achieve profit. Successful BI initiatives are possible with the support of intelligent technologies, tools and knowledge-based systems that are capable to sustain the introduced value chain.

REFERENCES

[1] Davenport, T.H. and Prusak, L. (2000) Working Knowledge: How organizations Manage What They Know, Harvard Business School Press, Boston.

[2] Gao, F, Li, M and Nakamori, Y (2002) Systems thinking on knowledge and its management: systems methodology for knowledge management, Journal of Knowledge Management, Vol.6, No.1, pp.7-17.

[3] Carvalho, R.B. and Ferreira, M.A.T. (2001) Using information technology to support knowledge conversion processes, Information Research, 7(1) [Available at http://InformationR.net/ir/7-1/paper118.html]

[4] Herschel, R.T., Nemati, H. and Steiger, D. (2001) Tacit to explicit knowledge conversion: knowledge exchange protocols, Journal of Knowledge Management, Vol. 5, No. 1, pp. 107-116.

[5] The Data Warehouse Institute (TDWI) web site, www.tdwi.org

[6] Negash S. Business intelligence. Communications of the Association for Information Systems2004; 13: 177-95.

[7] Gluchowski P. Business Intelligence - Konzepte, Technologien und Einsatzbereiche. HMD -Praxis der Wirtschaftsinformatik 2001; 222: 5-15 (in German).

[8] Bucher T, Dinter B. Process orientation of information logistics - An empirical analysis to assessbenefits, design factors, and realization approaches. 41th Annual Hawaii International Conference onSystem Sciences (HICSS-41); 2008: Waikoloa, Big Island, Hawaii: IEEE Computer Society; 2008.

[9] Hammer M, Champy J. Reengineering the corporation - A manifest for business revolution. NewYork: Harper Collins Publishers; 1993.

[10] Waterman D. A., A Guide to Expert Systems, Addison-Wisley, 1986.

[11] Kolodner, J. Case-based reasoning, Mogran Kaufmann Publishers Inc, 1993.
[12] Abdel-Badeeh M.Salem, Mohamed Roushdy and Rania A.HodHod ,
A Case Based Expert System For Supporting Diagnosis Of Heart Diseases,International Journal On Artificial Intelligence and Machine Learning, AIML, Tubungen, Germany, Volum 1, Dec. 2004, pp.33-39.
[13] Abdel-Badeeh M.Salem, Khaled A.Nagaty, Bassant M.El-Bagoury, A Hybrid Case-Based Adaptation Model For Thyroid Cancer Diagnosis, Proc. of 5th Int. Conf. on Enterprise Information Systems, ICEIS 2003 ,pp. 58-65 , Angres, France, 2003.
[14] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski. "Data Mining Methods for Knowledge Discovery", Kluwer Academic Publishers, 1998.
[15] I. H. Witten and E. Frank, Data Mining – Practical Machine Learning Tools and Techniques. 2nd ed Elsevier, 2005.
[16] A. M. salem, safia A. Mahmoud., "Mining patient Data Based on Rough Set Theory to Determine Thrombosis Disease", Proceedings of First Intelligence conference on Intelligent Computing and Information Systems, pp 291-296. ICICIS 2002, Cairo, Egypt, June 24-26,2002.
[17] Abdel-Badeeh M.Salem and Abeer M.Mahmoud, "A Hybrid Genetic Algorithm-Decision Tree Classifier", Proceedings of the 3rd International Conference on New Trends in Intelligent Information Processing and Web Mining, Zakopane, Poland, pp. 221-232, June 2-5, 2003.
[18] Muntean, M., Business Intelligence Approaches, WSEAS Conference on Mathematics and Computers in Business and Economics, Iaşi, 2012
[19] Porter, M. E., Competitive Strategy, Free Press, New York, 1980

Nadia Baeshen College of Business Administration, University of Business and Technology, Jeddah, Kingdom of Saudi Arabia, e-mail: nbaeshen@yahoo.com

Medical Images Understanding based on Computational Intelligent Techniques

Abdalslam AL-Romimah, Amr Badr, Ibrahim Farag

***Abstract*—**A computational intelligent system for regions of interest (ROIs) understanding is presented. It constitutes of fuzzy pulse-couple neural networks (FPCNNs) for ROIs and automatic understanding based on integer-CHC genetic algorithm (ICHCGA) with fuzzy artmap neural networks (FAMNNs). The system is applied on mammogram images, the mammogram understanding method consisting essentially of, automatic segmentation method based Fuzzy-PCNNs, and classification method based on ICHCGA feature selection and receiver operating characteristic (ROC) is generated by FAMNN for performance evaluation. The distinction between normal and abnormal cases by FAMNN is carried out by generated areas under ROC curve ranging from 0.88000% to 0.98604%, whereas distinction by MLPNNs is carried out by generated areas under ROC curve ranging from 0.72000% to 0. 86936%. FAMNN is used the distinction between benign and malignant mass is with fitness degree of 98.00% ranging from 0.87000% to 0.97845% under ROC curve, whereas distinction by MLPNNs with fitness degree of 92.00% ranging from 0.87000% to 0.95702% under ROC curve.

***Keywords*—**Digital Mammography, Fuzzy-PCNNs, FAMNN, Integer-CHC genetic algorithm.

I. INTRODUCTION

THIS field of bioinformatics is a crossway of numerous academic fields simultaneously; it provides an interface between medical sciences and information technology, using the most recent computerized means in analysis of medical images and thus diagnosing diseases and disorders on basis of establishing more analysis and understanding models for medical images.

Recently, in the last 30 years, there has been massive increase in the field of medical equipment and information technology in diagnostic imaging, the researchers have developed many different aid methods and systems in a field of biomedical informatics, whether traditional or computational intelligence techniques to improve tools of diagnostic effectiveness. For mammogram segmentation techniques [1] unsupervised and supervised approaches also known as model-based segmentation. Supervised approaches depend on prior information about image components if only objects or background. In unsupervised segmentation methods, image is partition into the set of regions dependent on specific features, intensity value, shape, texture and color.

Abdalslam .AL-Romimah is with the Department of Information Technology, Aden Community College & Al Saeed University , Yemen, Email: eng_rom32@yahoo:com.
A.Badr and I. Farag are with Faculty of Computers and Information, Cairo University, Egypt Email : see http :==www:fci:cu:edu:eg=StaffMembers:

S. Fu and J. K. Mui [2] divided unsupervised segmentation into three major groups: region-based methods, contour-based methods and clustering methods. Of course, all categorizations types for segmentation techniques are based upon color, intensity, or texture characteristics. Like Fu and Mu [3] considered the threshold methods as a special case of partitional clustering methods; where only two clusters are considered, threshold methods have been widely used for mass segmentation. There are two types of thresholding value which are used for image segmentation, hard and soft thresholding techniques, the hard thresholding techniques categorize in six groups as follows: histogram shape-based methods, clustering-based methods, entropy-based methods, object attribute-based methods, the spatial methods use higher-order probability distribution and local methods adapt the threshold value [4]. More recently, many studies for mammogram classification are presented; an automated mass detection method is presented by Timp et al [5] to detect temporal changes in mammographic masses between two consecutive screening rounds. Two kinds of temporal features, difference features and similarity features are designed to realize the interval change analysis. A SVM is employed as a classifier to detect the temporal changes in mammographic masses. The classification performance is evaluated with and without the use of temporal features. In experimental results, the database consisted of 465 temporal mammogram pairs containing 238 benign and 227 malignant cases. The Az = 0.74 without temporal features and 0.77 with the use of temporal features. Lcio et al [6] proposed an independent component analysis a feature extraction method and classification of mammograms with benign, malignant and normal tissues using three neural networks: MLPNN, probabilistic NN and radial basis function NN. The best performance is obtained with probabilistic NN, resulting in 97.3 success rate, with 100 of specificity and 96 of sensitivity. Retico et al [7] used the 16 features based on size and shape of the lesion are extracted: (area, perimeter, circularity, mean and the standard deviation of the normalized radial length, radial length entropy, zero crossing, maximum and the minimum axis of the lesion, mean and the standard deviation of the variation ratio, convexity, the mean, the standard deviation, the skewness and the kurtosis of the mass grey-level intensity values). For classification a standard three-layer feed-forward NN classifier merges the features into an estimated likelihood of malignancy. A data set of 226 massive lesions (109 malignant and 117 benign) is used. The system

performances are evaluated in terms of the ROC analysis, obtaining Az ranging 0.80 0.04 as the estimated Az. Pasquale et al [8] used the same 16 features extracted by [9] and the same dataset, but here feature selection procedure that are carried out on the basis of the feature discriminating power and of the linear correlations interplaying among them. 12 selected features out of the 16 computed use the Az of ROC evaluation with MLPNN. A MLPNN classifier is trained by error back-propagation algorithm. The masses dataset divided to 3 different categories: correctly, acceptably and non-acceptably segmented masses, Az=0.8050.030, 0.7870.024 and 0.7800.023, respectively. This paper is structured in four sections. In Section 2 automatic mammogram understanding is presented, it consisting essentially of, automatic segmentation based on FuzzyPCNNs, and automatic mammogram classification based on ICHCGA feature selection is performed FAMNNs categories classification. In section 3 FAMNN and MLPNNs evaluation results presented, and in section 4 the conclusions and future work.

II. AUTOMATIC MAMMOGRAM UNDERSTANDING

An automatic mammogram understanding method relates to improvements in image understanding methods, it consisting essentially of, automatic segmentation method based on fuzzy-pulse-couple neural networks (Fuzzy-PCNNs), and classification method based on integer-CHC genetic algorithm (ICHCGA) feature selection is performed with fuzzy artmap neural networks (FAMNNs) categories classification method.

A. AN AUTOMATIC SEGMENTATION

An automatic segmentation method based on Fuzzy-PCNNs method relates to improvements in image segmentation methods and systems. Fuzzy rule inference and fuzzy entropic threshold are adapted to improve a performance of PCNNs for image segmentation. Fuzzy entropic threshold is computed according to fuzzy max entropic that is depended on the image normalization, 2D image histogram and fuzzy partition, thus fuzzy entropic thresholding is adapted for PCNNs thresholding matrix ij [n]. The fuzzification and fuzzy rule are adapted to compute the coefficient matrix (i,j) (n) of a linking modulation layer of Fuzzy-PCNNs based fuzzy rule inference between the pixel and surround pixels in the image matrix, thus Fuzzy PCNNs method consisting essentially of, feeding and linking layers, Fuzzy-PCNNs filter based on the inverse of 2D Laplacian of Gaussian filter method of the resulted images after remove non-information regions, Fuzzy-PCNNs thresholding and Fuzzy-PCNNs pulse generator. Fuzzy-PCNNs filter is adapted as sharpening or high-pass filter, allow high frequencies pass and reduce the lower frequencies, and consequence is extremely sensitive to shut noise. To construct a high-pass filter, the kernel coefficients should be set positive near a center of kernel and in the outer periphery negative, for more details about Fuzzy-PCNNs see the equations from 2 to 6. The sequences of binary resulted images are filled by a polygon mask.

1. Fuzzy entropic threshold:

It is computed according to fuzzy max entropic [10] that is depended on the image normalization, 2D image histogram and fuzzy partition, thus it is adapted to get Fuzzy-PCNNs thresholding matrix ij [n].

2. Image normalization:

Normalization of the image resulted based on min-max normalization formula NZS see eq.[1], this image having gray levels ranging from lmin to lmax can be modeling as an array of fuzzy number; each element in the array is the value representing the degree of brightness of gray level between 0 and 1.

$$\mathrm{NZ}_S = \frac{\sum_{i=0}^{N} x(i,:) - l_{min}(i,:)}{l_{max} - l_{min}} \qquad (1)$$

3. 2D histogram :

2D histogram method proposed by Kirby and Rosenfeld [11] that is computed of resulted images, where each the bin of the 2D histogram represent a frequency of occurrence of each (level, local average gray level) pair. The bins form a surface with ideally two peaks corresponding to background and object regions. Thus, the pixels interior to the object or background are found mainly to the near-diagonal bins of a 2D histogram and off-diagonal bins being contributed by edges and noise in the region. For an n grey-level region there are obviously N^2 bins. By means of two thresholds S and T a 2D histogram is divided into 4 quadrants. Since the shaded quadrants of 2D histogram will in general contain information only about edges and noise that are ignored in the calculation. The quadrants 0 and 1 contain the distributions corresponding to the background and object classes.

4. Fuzzy partition:

In this section see [10], fuzzy entropy is adapted for image thresholding based on both intensity distribution and local information among pixels. The purpose of this method is to automatically determine the fuzzy region and optimal decay threshold parameter, therefore matrix of threshold ij , which is based on fuzzy entropy principle, given 2D histogram array of Lij region NM and K gray levels. The 2D histogram is divided to three regions: background region, fuzzy region and bright region. The background region is defined as the region with left top point (0, 0) and right button point (c, c). The overlapping region denoted by fuzzy region, which starts at point (a, a) and ends at point (c, c). For more details see the steps (3, 4, 5, and 6) in Fuzzy-PCNN method as it show below.

5. Fuzzy-PCNNs Model :

Firstly, fuzzy pulse-coupled neural networks (Fuzzy-CNNs) as developed model of PCNNs [12] shown as follows:

$$F_{ij}[n] = e^{-\alpha F \delta n} . F_{ij}[n-1] + s_{ij} + VF \sum_{kl} M_{ijkl} Y_{kl}[n-1] \qquad (2)$$

$$L_{ij}[n] = \sum_{kl} M_{ijkl} Y_{kl}[n-1] \quad (3)$$

$$U_{ij}[n] = F_{ij}[n].*(1+\beta_{ij}(n).* L_{ij}[n]) \quad (4)$$

Where β_{ij}=defuzzification (centroid method of β_{ij} as shown in fuzzy rules.

$$Y_{ij}[n] = \begin{cases} 1 & \text{if } U_{ij}[n] > \theta_{ij}[n-1] \\ 0 & \text{Othewise} \end{cases} \quad (5)$$

$$\theta_{ij}[n] = e^{-\alpha\theta\delta n}.\theta_{ij}[n-1] + V\theta Y_{ij}[n] \quad (6)$$

Where n=eq.11 in (see Fuzzy-PCNNs thresholding).

According to equations from 2 to 6 and 11, Fuzzy-PCNNs consisting essentially of, feeding and linking layers, Fuzzy-PCNNs filter based on the inverse of 2D Laplacian of Gaussian filter of the resulted images, Fuzzy-PCNNs thresholding and Fuzzy-PCNNs pulse generator. The main purpose of this method is separation the mammogram image regions well and full robotic. Fuzzy-PCNNs system consisting essentially of, 11 components shown as follows: Filter component is an inverse of 2D Laplacian of Gaussian filter of the resulted images, which is adapted as sharpening or high-pass filter, let high frequencies pass and reduce the lower frequencies, and consequence is extremely sensitive to shut noise. To construct a high-pass filter, the kernel coefficients should be set positive near a center of kernel and in the outer periphery negative.

Feeding and linking component, fuzzification component of linking coefficient (β), fuzzy rule inference component of linking coefficient, defuzzification component, linking modulation component of Fuzzy-PCNNs, Fuzzy-PCNNs thresholding component, which is computed according to fuzzy max entropic and Fuzzy-PCNNs pulse generator component, sequence of resulted images and polygon mask is adapted to fill the ROIs that are resulted from Fuzzy-PCNNs.

1. Fuzzy entropic thresholding:
As it shown in above.

2. Fuzzy-PCNNs filters:

A Fuzzy-PCNNs filter is an inverse of 2D Laplacian of Gaussian filter of the resulted images. Which is adapted as sharpening or high-pass filter, let high frequencies pass and reduce the lower frequencies, and consequence is extremely sensitive to shut noise. To construct a high-pass filter, the kernel coefficients should be set positive near a center of kernel and in the outer periphery negative.

3. Fuzzy-PCNNs feeding and linking:

This component represents the Fuzzy-PCNNs feeding and linking see eqs. (2, 3).

4. Fuzzification of linking coefficient β:

In this component a fuzzification of linking coefficient β is presented. The pixels in nn neighborhood region X surrounding each pixel (i, j) from feeding inputs, x(i, j) is gray level of (i, j) pixel in X. Let X (x(i, j)) denote the membership value represents the degree of coefficient between (i, j) pixel (Fij in PCNNs) with (Lij in PCNNs) in X. A fuzzy membership of region set X is mapping from X into interval [0,1]. For membership function, the homogeneity and edgeness measures are computed as fuzzy rules [13].

5. Fuzzy rule of linking coefficient:

The degree of membership for linking coefficient parameter β(i,j) is calculated as a matrix of values to knowing which the pixel Fij with surrounding neighborhood region Lij belongs to the four types (very_high, high, low, very_low).

The input space of the linguistic variable H(i,j) is comprised of the three fuzzy sets (low, med, high), and E(i,j) is comprised of two fuzzy sets labeled (low, high). Fuzzy rules can be defined as a conditional statement in the form:

If $(H_{i,j}$ is low) then $\beta_{i,j}$ is very_high (7)

If $(H_{i,j}\ is\ med)\ AND\ (E_{i,j}\ is\ high)\ OR$
$(H_{i,j}\ is\ high)\ AND\ (E_{i,j}\ is\ high)$ then $\beta_{i,j}$ is high (8)

If $H_{i,j}\ is\ med\ AND\ E_{i,j}\ is\ low$ then $\beta_{i,j}$ is low (9)

If $H_{i,j}\ is\ low\ AND\ E_{i,j}\ is\ low$ then $\beta_{i,j}$ is very_low (10)

Where $H_{i,j}$, $E_{i,j}$and $\beta_{i,j}$ are linguistic variables and {low, med, high}, {low, high} and {very_high, high, low, very_low} are linguistic values determined by fuzzy sets on the universes of discourse X and Y respectively. And fuzzy OR is defined as max (a, b) and fuzzy AND is defined as min (a, b).

6. Defuzzification of linking coefficient:

In this component, a defuzzification of fuzzy rule of linking [13], [14].

7. Fuzzy-PCNNs linking modulation

In this component, the linking modulation of Fuzzy-PCNNs is represented see eq.3.

8. Fuzzy-PCNNs thresholding:

In this component, in Fuzzy-PCNNs, an optimal decay thresholding parameter αθδn is calculated as follows:

$$\alpha\theta\delta n = \max\left((t_{tiss} + \mu_{maxfn}), max(\mu_{max})\right) \quad (11)$$

Where μ_{maxfn} is the global maximum fuzzy entropy [15] of a normalize image (maximum entropy of fuzzy number) in fuzzy partition, t_{tiss} is the coefficient based on the type of

mammogram tissue, which is determined based on the experiments. And the max (μ_{max})) is the maximum fuzzy entropy of μ_{max} matrix. Where μ_{max} is a matrix of membership for optimal thresholding (maximum fuzzy entropy) for feeding $F_{i,j}$ with its surrounding neighborhood $L_{i,j}$.

9. Fuzzy-PCNNs pulse generator: See eq.5.
10. Resulting images:

In this component, a sequence of binary resulted images. Seeeq.5.

11. Filling of ROIs:

In component polygon mask is adapted to fill the ROIs that are resulted from Fuzzy- PCNNs system. In binary image, the regions of interest are detected by polygon mask (tracing boundary contours) and fill it using filtering the ROI from original image, which returns an image that consists intensity values for pixels in locations where ROI image contains 1's, and unfiltered values for pixels in locations where ROI image contains 0's. Then save the each region of interest as image.

Fuzzy-PCNNs (Pseudo-Code) :

Fuzzy-PCNNs system steps for mammogram mass segmentation and micro-calcification detection passes through various components as shown in follows:

1. Before use Fuzzy-PCNNs for mammogram mass segmentation and micro-calcification detection, a system of automatic tissue types identification is worked, a MLPNNs classifier is adapted to know the tissue type. If tissue type is not a dense tissue, in the other hand, one from first four types, a mammogram image is enhanced by AHE method on a specific range of gray levels.
2. Image normalization.
3. 2D histogram is calculated one only.
4. Set initial values of all Fuzzy-PCNNs parameters and matrixes.
5. βi, j(n) is calculated at each iteration. The coefficient parameter βi,j(n) is different from Fi,j with its surrounding pixels in the same region to other. Thus the coefficient degree is determined based on the strong relationship between the pixels with its surrounding pixels.
6. Given an eq. 11 an optimal decay thresholding parameter αθδn is obtained base on a maximum of local maximum fuzzy entropy matrix μ_{max}, the global maximum fuzzy entropy of the normalize image (maximum entropy of fuzzy number μ_{maxfn}), and a value of t_{tiss} parameter, which is based on MLPNNs result.
7. The binary images are resulted using the Fuzzy-PCNNs. These images are included the ROIs (exactly a first binary image). All the ROIs are detected by polygon mask to draw each ROI separately. Therefore a new binary image for each ROI is created separately based on values of its boundary (by tracking boundary is aforementioned). Each boundary of a ROI has same the location in original image. The ROI is filled using a filtering it with original image.

B. AUTOMATIC MAMMOGRAM CLASSIFICATION

Automatic mammogram classification relates to improvements in computational intelligent methods for classification of medical images. This method consisting essentially of, feature extraction, feature selection and classification.

1. Feature extraction:

In this section, various special methods are adapted to extract the ROIs features and generate a features matrix. Textural features such first order statistics, second order statistics features of well-known gray level co-occurrence matrixes (GLCMs), gray level run length matrixes (GLRLMs) features, fractional dimension features and multilevel wavelet decomposition features. Shape features and density features also are extracted.

2. Feature selection :

Mammogram feature selection relates to improvements in computational intelligent methods for the medical images classification. Integer-CHC Genetic Algorithm (ICHCGA) is proposed to attain a best balance between the exploration and exploitation [16], CHC (cross-generational elitist selection, heterogeneous recombination, and cataclysmic mutation). This is accomplished by maintaining diversity in the population and allowing the algorithm to focus in several areas of search space simultaneously, and it is used to force

diversity onto a population. A CHC algorithm is developed to solve the problems of premature convergence that genetic algorithm frequently suffers, and it uses a conservative strategy of selection. In ICHCGA, integer-coded is adapted in lieu of binary coded, because the last one require a de-codification step to apply the fitness function and also does not fit well when the number of features is fixed.

2.1. Integer coded :

For feature subset selection integer coded is not require a de-codification step to apply the fitness function and does fit well when the number of features is fixed.

2.2. Fitness function :

Fitness(subset) = |(accuracy/F AMNN error-rate)| 12)

2.3. CHC method:

ICHCGA based on CHC (cross-generational elitist selection, heterogeneous recombination, and cataclysmic mutation) is adapted to force diversity onto a population, when it may have become trapped around a sub-optimal solution .

a) Elitist selection:

This method is one of the elitist steady-state selection algorithms, which explicitly borrow from the (+) evolutionary strategies [17], [18], [19]. It is based on survival of fittest instead of reproduction with emphasis; the survivors are chosen from the old parent population to the next generations parent population and select the remaining members from the offspring population. And the survivors are the elite chromosomes having the best criterion value determined by the fitness function.

b) Incest Prevention:

To avoid premature convergence, ICHCGA employs the incest prevention mechanism, which can be used to promote exploration at the start of the search. If the minimum difference between parents is relatively large, the offspring will be sufficiently different to promote exploration. As this required difference decreases in later generations, the similarity of the parents and therefore the offspring increases and this focuses the search into a particular region of search space. In other words, two parents are only mated, if their Hamming distance (in binary coded) is above a threshold and an increase in the mutation probability is not required. Therefore, before applying HUX to two parents, the dissimilarly between them is measured by a Hamming distance of the gene strings, which is a count of number of the differing bits. In case, the integer or real coding, the dissimilarly is obtained by sum of the absolute differences between the values across all loci (their Manhattan distance or Euclidean distance). The individuals are able (or allowed to them) to mate and produce offspring, only if, the average of Euclidean distance is above of a certain (mating threshold) is achieved. This mean the elite chromosomes are rank-ordered from top down only chose points that have a decoded, Euclidean distance greater than a threshold from all previously selected points. Only these points are used in mating and the parents and offspring are used to cast out several new offspring.

c) HUX crossover:

Using HUX, the substrings are switched between offspring with a probability Pc ross; and this probability decreases with each generation. In essence this is a biased uniform crossover between integer-coded strings (where the fitness of the parent determines the probability that its gene will be expressed), and the bias increases with each generation. Although the elite chromosome was paired with another parent chromosome it remained unchanged after crossover was performed.

d) cataclysmic mutation:

To keep on the production of offspring with maintains diversity and slows population convergence a cataclysmic mutation (called re-start mechanism) is applied. According to CHC adaptive algorithm, the value of this cataclysm threshold is decreased as the population converges and individuals become more similar. This threshold is calculated as follows:

The initial threshold is set at L/4, where L is the length of the chromosomes, or threshold:= $MP * (1.0 - MP) * L$, where MP is mutation probability (0.35). According to Eshelman scheme [16], the cataclysmic mutation can also be used to construct families if it is extended to use multiple parents with a given threshold separation instead of simply using just the elite chromosomes. It must be emphasized that this process generates multiple

offspring from a single parent by only using a mutation operator. If no offspring are inserted into the new population at the next generation, then the threshold is reduced by one. In other words, In order to avoid very slow convergence, threshold will be also decremented by one, when no improvement is achieved respect to the best chromosome of the previous generation. On the other hand, whenever the population converges towards a certain points, a cataclysm occurs (If the threshold¡0).

e) restart (can be included in cataclysmic mutation):

The new population includes one copy of the best individuals, while the rest of the population is generated by mutating some percentage of genes of such best individuals. In other way, the elite chromosomes are used as a template to re-seed the population. Randomly, the rate changing of bits is 35 in the template chromosome to form each of the other chromosomes in the population. The Euclidean threshold is reset and the algorithm resumes in the usual manner.

3. Classification :

Fuzzy artmap neural network (FAMNN) with receiver operating characteristics (ROC), FAMNN for training and testing, and ROC for evaluate the performance of FAMNN. FAMNN is one of the incremental learning algorithms are presented by Carpenter et al [20], [21], [22], in response to stability-plasticity dilemma (the catastrophic forgetting phenomenon through neural network learning). This technique is characterized by the following:

1) The FAMNN (nonlinear separability): able to build decision boundaries that separate classes of any shape and size.
2) The FAMNN (overlapping classes): creates decision boundaries to minimize the misclassification for all overlapping classes. In other words, there is no overlap between hyperboxes of different classes.
3) The FAMNN (training time): needs only one pass to learn and refine its decision boundaries.

III. FAMNN AND MLPNNS EVALUATION RESULTS

In this work, the FAMNN and MLPNNs performance are evaluated by fitness (an accuracy or error rate) of ICHCGA and AUC of ROC.

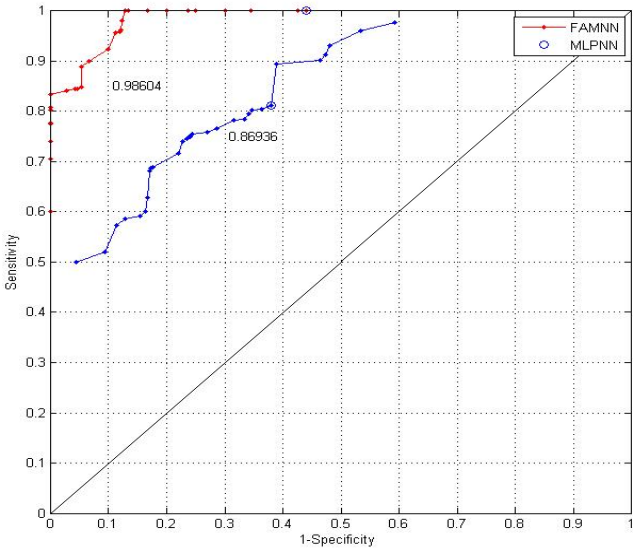


Fig. 1. : ROC curves for normal and abnormal tissues for comparison between FAMNN- ICHCGA and MLPNNs performance.

1) The discernment results between two classes using FAMNN and MLPNNs shown as follows:

a) For normal or abnormal see (Table I, Table II) and (Figure1). And the AUC of ROC in the best population using FAMNN is higher than MLPNNs.

b) For benign or malignant see (Table I; Table II) and (Figure2). And the AUC of ROC in the best population using FAMNN is higher than MLPNNs.

2) The discernment results between multi-class using FAMNN and MLPNNs shown as follows:

a) For normal or benign or malignant see(Table I, Table II) and (Figure 3).And the AUC of ROC in the best population using FAMNN is higher than MLPNNs. Finally, we note that best results are at using GA - FAMNN for stepwise GA-MLPNNs see Table II ,Table II .

IV. CONCLUSIONS AND FUTURE WORK

The main goal of this thesis has addressed the investigation of computational intelligence techniques and their applications especially in medical images understanding. Using MIAS dataset, 200 mammograms are used for mass segmentation and classification, 96 mammograms have a normal case with all tissue types (fatty, glandular, density),

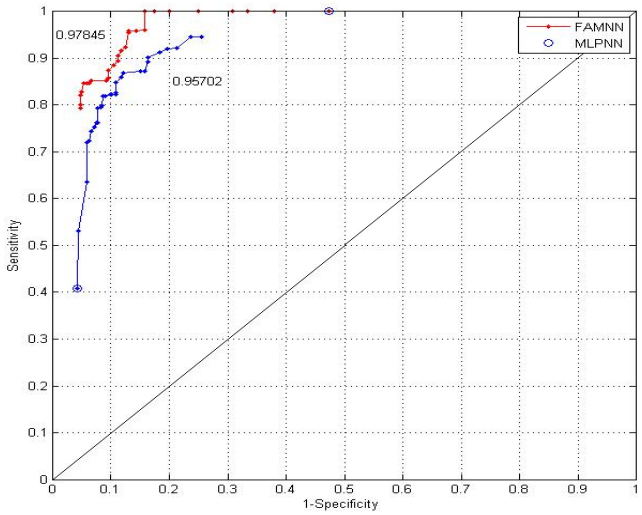


Fig. 2. : ROC curves for benign and malignant tissues for comparison between FAMNN- ICHCGA and MLPNNs performance.

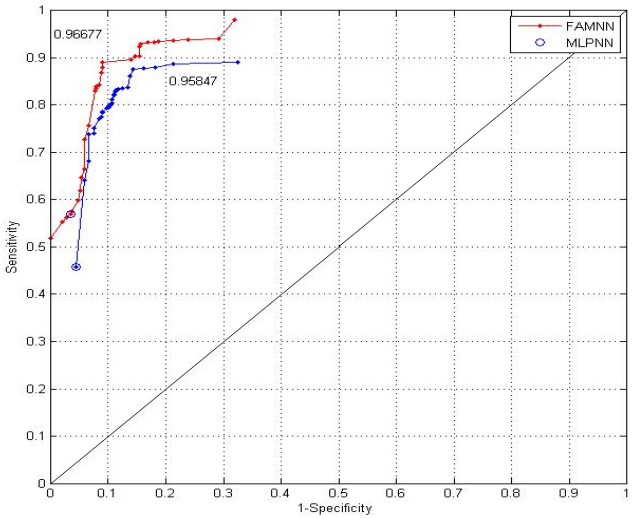


Fig. 3. : ROC curves for normal, benign and malignant tissues for comparison between FAMNN- ICHCGA and MLPNNs performance.

53 mammograms have a benign mass with all tissue types and 41 mammograms have a malignant mass with all tissue types. Other 18 mammograms are used for micro-calcification detection, 11 mammograms have a benign case from all tissue types (fatty, glandular, density) and 7 mammograms have a malignant case from all tissue types. The potential of a novel segmentation technique based on Fuzzy-PCNNs is investigated. This computational intelligence model is unsupervised, context sensitive, robotically and invariant to a tissue type. Therefore, Fuzzy-PCNNs own rather interesting properties for the automatic processing of most applications. The Fuzzy-PCNNs approach aiming at separate the ROIs in the image (high spots) based on fuzzy membership degree of coefficient between pixel and it neighbors and fuzzy membership degree of different between them by maximum fuzzy entropy (soft thresholding) rather than segment the ROIs based on initial value (hard thresholding). For feature extraction 188 features are used, whether statistical or geometric, as well as Wavelet technique is used in order to deal with the ROI with multi-scale. Also for feature selection, the ICHCGA is used to select the best available features. For discernment between normal and abnormal, 50 rows from normal data and 50 rows from abnormal data are used as training set. And 46 rows from normal data and 44 rows from abnormal data are used as testing set. The best results when FAMNN is used compare with MLPNNs are as follows:

error rate = 0.0000, AUC of ROC =0.98604, features selection number=6. For discernment between benign and malignant, 28 rows from benign data and 21 rows from malignant data are used as training set. And 25 rows from benign data and 20 rows from malignant data are used as testing set. The best results when FAMNN is used compare with MLPNNs are as follows: error rate = 0.0200, AUC of ROC =0.97845, features selection number=6. For discernment between normal, benign and malignant, 50 rows from normal data, 28 rows from benign data and 21 rows from malignant data are used as training set. And 46 rows from normal data, 25 rows from benign data and 20 rows from malignant data are used as testing set. The best results when FAMNN is used compare with MLPNNs are as follows: error rate = 0.0111, AUC of ROC =0.96677, a features selection number=6.

- Future Work : The work in Future will be to develop a medical images understanding for diseases prognosis. This model will use to help discover possible cancers before its occurring in the future based on a time series of mammogram images for women, who come early to screen up. With other view, a Fuzzy-PCNNs model can be used to develop other applications such as automatic change detection in very high resolution images (satellite images analysis).

ISBN: 978-1-61804-236-1

TABL I
I NPUT AND RESULT OF BREAST CANCER CLASSIFICATION BY FAMNNS AND EVALUATED IT PERFORMANCE USING ICHCGA AND AUC OF ROC CURVE.

Classification according to breast cancer tissue	Training set	Testing set	Fitness	Az ROC	No. Features selection	Generation
Normal abnormal	50-50	46-44	1.0000	0.98604	6	240
Benign-malignant	28-21	25-20	0.9800	0.97845	6	270
Normal-benign-malignant	50-28-21	46-25-20	0.9889	0.96677	6	290

TABLE II
INPUT AND RESULT OF BREAST CANCER CLASSIFICATION BY MLPNNS AND EVALUATED IT PERFORMANCE USING ICHCGA AND AUC OF ROC CURVE.

Classification according to breast cancer tissue	Training set	Testing set	Fitness	Az ROC	No. Features selection	Generation
Normal abnormal	50-50	46-44	0.9578	0.98604	6	240
Benign-malignant	28-21	25-20	0.9200	0.97845	6	270
Normal-benign-malignant	50-28-21	46-25-20	0.9100	0.96677	6	290

REFERENCES

[1]. Arnau Oliver, Jordi Freixenet, Joan Marti, Elsa Perez, Josep Pont, Erika R.E. Denton,Reyer Zwiggelaar.A review of automatic mass detection and segmentation in mammographic images. Medical Image Analysis 14 (2010) 87110.

[2]. K. S. Fu and J. K. Mui. A survey on image segmentation. Pattern Recognition,13:316, 1981.

[3]. Radhika Sivaramakrishna, Nancy A. Obuchowski, William A. Chilcote, Kimerly A. Powell. "Automatic Segmentation of Mammographic Density".academic radiology, Volume 8, Issue 3, Pages 250-256 (March 2001) .

[4]. Mehmet Sezgin. Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging 13(1), 146165 (January 2004).

[5]. S. Timp and N. Karssemeijer. A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography.IEEE Transactions on Medical Imaging, 31(5):958971, 2004.

[6]. Lcio F.A. Campos, Aristfanes C. Silva, and Allan Kardec Barros.Diagnosis of Breast Cancer in Digital Mammograms Using Independent Component Analysis and Neural Networks. CIARP 2005, LNCS 3773, pp. 460 469, 2005.

[7]. Retico, P. Delogu, M.E. Fantacci, P. Kasaec.An automatic system to discriminate malignant from benign massive lesions on mammograms. Nuclear Instruments and Methods in Physics Research A 569 (2006) 596600.

[8]. Pasquale Delogu, Maria Evelina Fantacci, Parnian Kasae, Alessandra Retico ,Characterization of mammographic masses using a gradient-based segmentation algorithm and a neural classifier. Computers in Biology and Medicine 37 (2007) 1479 1491.

[9]. Retico, P. Delogu, M.E. Fantacci, P. Kasaec.An automatic system to discriminate malignant from benign massive lesions on mammograms.Nuclear Instruments and Methods in Physics Research A 569 (2006) 596600.

[10]. H. D. Cheng, Yen-Hung Chen, Fuzzy partition of two-dimensional histogram and its application to thresholding, Pattern Recognition, Volume 32, Issue 5, May 1999.

[11]. R.L. Kirby and A. Rosenfeld, A note on the use of (gray-level, local average gray-level) space as an aid in threshold selection. IEEE Trans. Syst. Man Cybernet. SMC-912 (1979), pp. 860866.

[12]. Lindblad, Th.; and Kinser, J.M. (1998). Image Processing using Pulse- Coupled Neural Networks, Perspectives In Neural Computing. Springer-Verlag Limited. ISBN 3-540-76264-.

[13]. Zadeh, L.A. Outline of a New Approach to the Analysis of Complex Systems and Decision Processes . Information science, Vol.9, pp.43-80. (1973).

[14]. B. Riecan, D. Markechova, The entropy of fuzzy dynamical systems, general scheme and generators, Fuzzy

Sets and Systems, Volume 96, Issue 2, 1 June 1998, Pages 191-199.

[15]. H. D. Cheng, Jim-Rong Chen, Automatically determine the membership function based on the maximum entropy principle Information Sciences, Volume 96, Issues 3-4, February 1997, Pages 163-182.

[16]. L. Eshelman, The CHC Adaptive Search Algorithm, How to Have Safe Search When Engaging in Non-traditional Genetic Recombination,Morgan Kaufman, S. 265 283-1991.

[17]. Rechenberg, Evolutions strategies, From mannHolzboog, 1973.

[18]. H. Schwefel, Numerical optimization of computer models (M. Finnis, trans.), Chichester: John Wiley, 1981 (Original work published 1977).

[19]. Larry J. Eshelman, James D. Schaffer, Method for optimizing the configuration of a pick and place machine, United States Patent No. 5,390,283, 14 February 1995.

[20]. R. Polikar, L. Udpa, S. Udpa, V. Honavar, Learn++: An incremental learning algorithm for multilayer perceptrons. Proceedings of 25th. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 6, pp: 3414-3417, Istanbul, Turkey,2000.

[21]. [3] R. Polikar, L. Udpa, S. Udpa, V. Honavar. Learn++: An incremental learning algorithm for supervised neural networks. IEEE Transactions on Systems, Man, and Cybernetics.Part C: Applications and Reviews, Vol. 31, No. 4, pp: 497-508, 2001.

[22]. S. Grossberg, Adaptive pattern recognition and universal encoding II:Feedback, expectation, olfaction, and illusions, Biol. Cybern., vol. 23,pp. 187202, 1976.

Efficient Answering of XML Queries using Holistic Twig Pattern Matching

Divya Rajagopal and Dr. Miraclin Joyce Pamila J. C.

***Abstract*—**XML is used as a standard for expressing semi structured data, a form of data that does not conform to the formal structure of relational data models but nonetheless contains tags to separate semantic elements and enforce hierarchies of records and fields within the data. Semi structured data model allows information from several sources with related but different properties to be integrated, thereby enabling sharing of information. Improving the efficiency of answering the queries issued on such data is therefore a great challenge. Twig pattern matching is a critical operation for XML query answering and holistic approaches have shown superior performance over other methods. In this paper, we propose a novel holistic twig pattern matching algorithm which performs optimal matching for twigs with both AD (Ancestor Descendent) edges as well as PC (Parent Child) edges, while prior algorithms claim optimality for twigs with only AD edges.

***Keywords*—**ancestor-descendant, holistic, parent-child, XML

I. INTRODUCTION

An XML document consists of data enclosed within a set of user-defined tags. The tags should be properly nested, they should be paired, and there should be one and only one root tag and so on. XML offers simplicity, flexibility, standardization and interoperability. Hence XML is being widely used as a data representation format for representing nearly all kinds of data. However XML documents are often very large and have a deeply nested structure. And also the XML data can be very complex. Hence efficient pattern matching algorithms are needed to retrieve the data from the XML documents by answering the queries given on the documents.

A query on the XML document describes a tree-shaped or hierarchical search pattern, which is often referred to as a twig pattern [2]. XML queries are thus called tree queries or twigs and the relationships (AD or PC) between the components of the twig are represented as edges. Single backslash (/) is used to represent a parent-child edge or PC edge. When / is used at the beginning of a query, for example /book, it will define an absolute path to node "book" relative to the root. In this case, it will only find "book" nodes at the root of the XML tree. When / is used in the middle of a query, for e.g. /book/author, it will define a path to node "author" that is a direct descendant (i.e. a child) of node "book". Double backslash (//) is used to represent an ancestor-descendant edge or AD edge. When // is used at the beginning of a query, for example //book, it will define a path to node "book" anywhere within the XML document. In this case, it will find "book" nodes located at any depth within the XML tree. When // is used in the middle of a query, for e.g. /book//author, it will define a path to node "author" that is any descendant of node "book".

The core operation of XML query answering is twig pattern matching: finding in an XML document tree 'D', all matches of a given tree-type query 'Q' called twig. A match is identified by a mapping from nodes in 'Q' to nodes in 'D' such that query node predicates are satisfied by the corresponding document tree nodes and also the structural relationships (AD or PC) between query nodes are satisfied by the corresponding document tree nodes.

The answer to query 'Q' with 'n' nodes can be represented as an n-ary relation where each tuple (d1,…,dn) consists of the document tree nodes that identify a distinct match of 'Q' in 'D'.

Holistic twig pattern matching approaches avoid large sets of irrelevant intermediate results by considering the structural inter-dependencies among the XML elements. Holistic approaches optimize pattern matching in two phases:

1. Labeling: assigning to each node x in the data tree t, an integer label label(x) that captures the structure of t.
2. Computing: exploiting the labels to match a twig pattern p against t without traversing t again.

In this paper, we propose a novel holistic twig pattern matching algorithm which performs optimal matching [2] for twigs with both AD (Ancestor-Descendent) edges as well as PC (Parent-Child) edges.

II. EXISTING SYSTEMS

TwigStack [2] is a holistic twig join algorithm that ensures that no large intermediate results are produced. When the query has only ancestor-descendant relationships between the elements, TwigStack is I/O and CPU optimal but it is suboptimal when the query has parent-child relationship among the elements. It is suboptimal as the overall computation cost for a twig pattern is proportional not just to the sizes of the input and the final output but also to the sizes of the intermediate results. GTwigMerge [3], a basic framework for holistic processing of AND/OR-twigs works correctly when AND/OR twig queries contain parent-child QNodes. However, the optimality in terms of worst-case I/O and CPU cost is no longer guaranteed. There are two reasons for the sub-optimality. First, if some output nodes are parent-child QNodes, a path solution may turn out not to join with any other path solutions. Thus, irrelevant I/O access is caused. Second, if some OR-predicates in an AND/OR-twig contain

parent-child QNodes, a path solution may contain an element node that eventually turns out not to satisfy all its OR-predicates. TwigStackList [4] is another holistic twig join algorithm which is I/O optimal for queries with only ancestor-descendant relationships below branching nodes. The optimality cannot be proved for the case where parent-child relationships appear only in edges below non-branching nodes. TwigStackList¬ [5] is a new algorithm to match NOT-twig queries holistically. In a NOT-twig, this algorithm can guarantee the I/O optimality only when all the positive edges below branching nodes are ancestor-descendant relationships.

III. TREE REPRESENTATION OF XML AND TWIG

As the XML and the twig are hierarchical, they are represented using a tree data structure. A sample XML tree representation is shown in Fig. 1. Each node in the tree corresponds to an XML element. The root node corresponds to the root element, the intermediate nodes to sub elements, the leaf node to values. Each edge corresponds to an element-sub element or element-value relationship.Each non-leaf node in the XML tree can have multiple, variable number of children. Hence instead of a linked list implementation of the tree, a more optimized tree representation (shown in Fig. 2) is used. In this representation, each non-leaf node has two pointers: a pointer to the first child and a pointer to the next sibling. The optimized XML tree representation of the tree used in Fig. 1 is demonstrated in Fig. 3. Because each node has at most only two children, the new tree is a binary representation of the previous tree.

The twig considered in this paper is a plain twig which contains only a single path from root to leaf. Sample query trees are shown in Fig. 4. Every node in the twig, called a QNode or query node, associates to an element type or tag name in a tree database. For programmatic purposes, a QNode records its location step axis or edge type as either "//" or "/" for edge test, and a tag name for node test. Therefore, the content of a QNode takes the general format of "/tag" or "//tag."

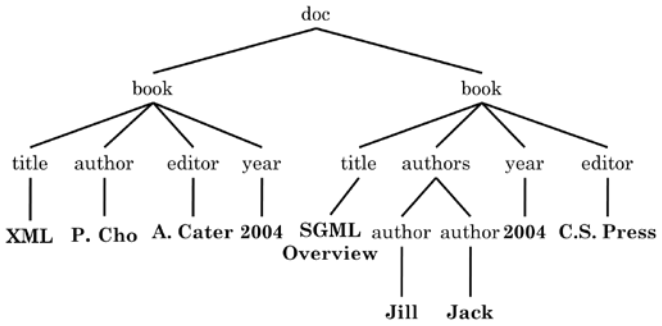


Fig. 1. Representation of XML tree

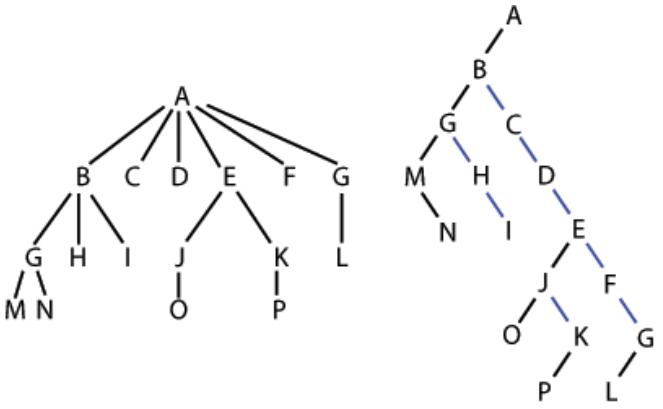


Fig. 2. Data structure for XML tree representation

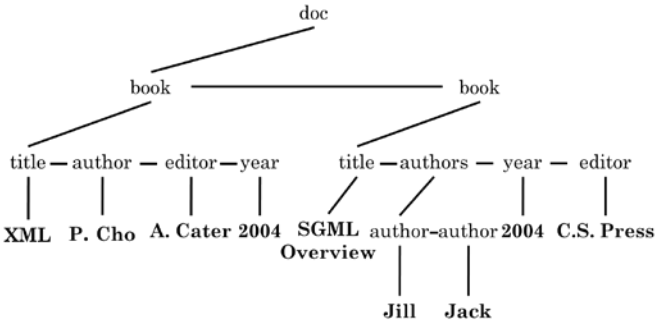


Fig. 3. Optimized representation of XML tree

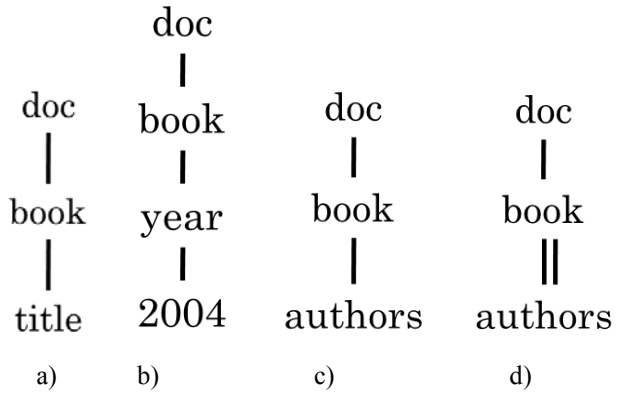


Fig. 4. Representation of XML query trees

IV. XML TREE LABELING

The aim of data tree labeling schemes is to determine the relationship (i.e., Parent-Child or Ancestor-Descendant) between two nodes of a tree from their labels alone. Each node in the XML tree is given a unique identity called label or region code. In this paper, the triplet region encoding scheme which is obtained through pre-order traversal of the document tree is used. Each label consists of three parts: start position, end position, level. The encoded version of the XML tree shown in Fig. 1 is shown in Fig. 5.

ISBN: 978-1-61804-236-1

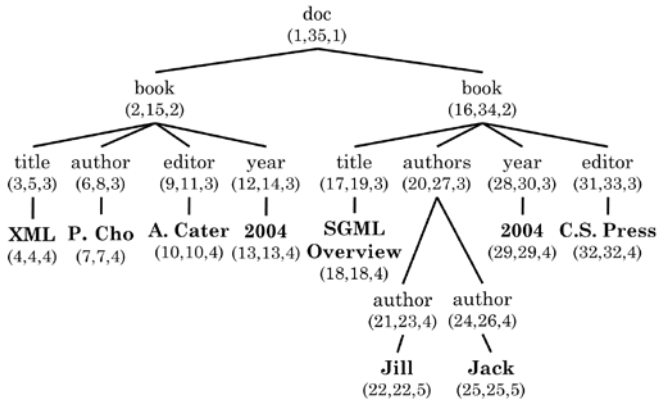


Fig. 5. Encoded XML tree representation

The relative positional information obtained is as follows: Let x and x' be two nodes labeled (S, E, L) and (S', E', L'), respectively. Then,

- x' is a descendant of x if and only if S' > S and E > E'. Thus the edge between x and x' represents an ancestor-descendant edge.
- x' is a child of x if and only if S' > S and E > E' and L'=L+1. Thus the edge between x and x' represents a parent-child edge.

V. TWIG PATTERN MATCHING MAIN ALGORITHMS

Fig. 6 presents the main algorithm "TwigMerge" of the second phase called the computing phase of the proposed twig pattern matching process. TwigMerge uses the labels to compute the answers to the twig. All the supporting functions are as described in [1].

```
ALGORITHM TwigMerge(root)
1: while not end(root) do
2:     q = GetQNode(root);
3:  if q == null then
4:      continue;
5:  if not isRoot(q) then
6:         cleanStack(S_QParent(q),C_q);
7:     cleanStack(S_q, C_q);
8: if isRoot(q) or (not empty(S_QParent(q))) then
9:         if not isLeaf(q) then
10:            push(S_q, C_q,isRoot(q)? -1 : top(S_QParent(q)));
11:        else
12:            outputPathSolutions(C_q);
13:   C_q ->advance();
14: end while
```

Fig. 6. TwigMerge holistic twig pattern matching algorithm

Some important features of TwigMerge are highlighted as follows:

1. TwigMerge receives (from GetQNode shown in Fig. 7) either a valid QNode q or an invalid QNode, denoted by null (the validness is checked at line 3). An invalid QNode is generated when a non-top level recursive call into this function fails to find a QNode associated with a fully qualified element. But since noncontributing elements have been skipped, the main algorithm quick jumps to its next iteration (at line 4) to start a new call to GetQNode for getting the next valid QNode.
2. No stacks are allocated for non-output QNodes nor for any leaves since the contributing elements which correspond to the leaf can be directly grabbed from the associated stream.
3. All critical processing logics are encapsulated in the key supporting function, GetQNode and other lower level supporting functions edgeTest and hasExtension. In addition to feeding the main algorithm with the next query node to be processed, GetQNode also checks the candidacy of the elements in the input streams and guarantees that for the next QNode returned to the main algorithm the current element in the associated stream is fully qualified. Therefore, TwigMerge achieves both I/O and CPU optimality not only with AD edges but also with PC edges.
4. "Stack cleaning" (lines 6 and 7) is needed in TwigMerge solely because each time after outputting path solutions, some elements on the stacks may become irrelevant for future path solutions and must be cleaned out. In most prior algorithms such as TwigStack, stack cleaning is required to get rid of those noncontributing elements that may have been tentatively added to the stacks but are actually noncontributing.
5. TwigMerge moves the element associated to q from stream to stack if it is not a leaf (at lines 8 to 10), otherwise (q is a leaf) outputs the path solutions currently on the stacks (lines 9 and 10).

Another critical supporting function, hasExtension (refer Fig. 8), implements our definition of a match for a twig. It helps in ensuring that only relevant contributing nodes are taken for processing. For leaf nodes, it performs edgeTest to confirm the relevancy with respect to edge-type (AD or PC) and for non-leaf nodes, it performs testing of further extensions.

```
FUNCTION GetQNode(q)
1: if isLeaf(q) then
2:     return q;

3: for each q_i ∈ Qchildren(q) do
4:     q_0 = GetQNode(q_i);
5:  if q_0 != q_i then
6:         return q_0;
7: end for
8: q_max = getMaxQChild(q);
9: while C_q ->end < C_qmax ->start do
10:   C_q ->advance();
11: end while

12: q_min = argmin_qi { C_qi ->start}, q_i ∈Qchildren(q);
13: while C_q ->start < C_qmin ->start do
14:   if hasExtension(q) then
```

```
15:      return q;
16:   else
17:      Cq ->advance();
18: end while
19: if hasExtension(qmin) then
20:      return qmin;
21:   else
22:      Cqmin ->advance();
23. if end(q) then
24:   return null;
25: else
26:   return GetQNode(q);
```

Fig. 7. Pseudocode for GetQNode function

```
FUNCTION hasExtension(q)

1: for each qi ∈children(q) do
2:    if isLeaf(qi) then
3:       return edgeTest(Cq, qi)
4:    else
5:       return (edgeTest(Cq, qi) and hasExtension(qi))
6: end for
```

Fig. 8. Pseudocode for hasExtension function

The function edgeTest is presented in Fig. 9. The while loop (at lines 8 and 9) brings an important optimization: fast skipping noncontributing elements in stream T_q until the cursor moves over the range of the parent element e. Holistic twig joins typically disallow backtracking of stream cursors to guarantee linear time complexity.

```
FUNCTION edgeTest(e,q)
1: while not end(Cq) do
2:    if e.start < Cq ->start and e.end > Cq ->end then
3:       if q.edgeType == '//' then
4:          return true
5:    else if e.level == Cq ->level-1 then
7:       return true
8: if Cq ->end < e.end then
9:       Cq ->advance()
10:   else
11:      break
12: end while
13: return false
```

Fig. 9. Pseudocode for edgeTest function

VI. COMPLEXITY OF TWIGMERGE

Given a twig query Q, the parameters used are:

- |Input| stands for the total size of all the input streams relevant to query Q
- |Output| stands for the total count of the data elements included in all output twig instances produced for query Q

The I/O cost of TwigMerge consists of three parts: the I/O cost for accessing all the relevant input stream elements and the I/O cost for dealing with the intermediate path solutions plus the I/O cost for outputting the final twig solutions. Since in TwigMerge, the stream cursors are always advanced and never backtracked, the first part of the I/O cost is the total size of all relevant input streams. For the second part, since TwigMerge is optimal with both AD and PC edges—i.e., it never produces useless intermediate path solutions, the I/O cost of this part is two times (for first output and then input) of the total final output size, i.e., 2 * |Output|. And the third part (for outputting the final results), of course, is |Output|. The total I/O cost for TwigMerge is the sum of the above three parts = |Input| + 3 * |Output|

The CPU cost analysis for TwigMerge is analogous. The CPU cost also consists of three parts. The first part is the time spent on computing the path solutions, the second part is the time spent on dealing with the obtained intermediate path solutions (output, input, and merging), and the third part is on outputting the final twig solutions. The main structure of TwigMerge is a loop that repeats no more than |Input| times, which is the total number of elements in all the input streams because noncontributing elements are skipped at line 10, 17, and 22 of GetQNode (refer Fig. 7) or by the optimization rendered by the primitive function edgeTest (refer Fig. 9). So the first part of the CPU cost is linear to the input size. The second part depends on how many intermediate path solutions are produced and how many of them are going to be merged to form the final output twig solutions. As TwigMerge does not produce any unused intermediate path solutions (it actually does not push any noncontributing elements onto any stack), the second part of the cost is linear to and solely decided by the output size |Output|. And the third part of course is also linear to the output size. Added together, for the overall CPU cost of TwigMerge, exactly the same result as that derived for the I/O cost is obtained (cost equations omitted). The above cost analysis results shows that TwigMerge has both optimal I/O cost and optimal CPU cost for twigs with both AD and PC edges.

VII. RESULTS

To avoid potential bias of using a single data set, two XML data sets are used for this study. The first is a docBook data set which contains the details of various books. The second data set is the TreeBank data set, downloaded from the University of Washington XML Repository website [6]. The XML document of the TreeBank data set is deep and has many recursions in structure. This data set takes 82 MB memory, consisting of 2.4 million data nodes. The average depth of TreeBank is 7.8 and the max depth is 36. Fig. 10 shows the parsing of docBook XML document, construction of the XML tree. The number of leaf nodes, non-leaf nodes, total number of nodes and maximum depth are also shown. The twig results for the queries on the docBook XML Document are shown in Fig. 11, Fig. 12. The twigs contain combinations of ancestor-descendant and parent-child edges. The corresponding execution time (in milliseconds) is also displayed. Fig. 13 shows the parsing of treebank XML document, construction of and labelling of the XML tree. The number of leaf nodes, non-leaf nodes, total number of nodes and maximum depth are also shown. The twig results for the query on the treebank XML Document is shown in Fig. 14. The twig contains an ancestor-

descendant edge. The corresponding execution time (in milliseconds) is found to be 10395.

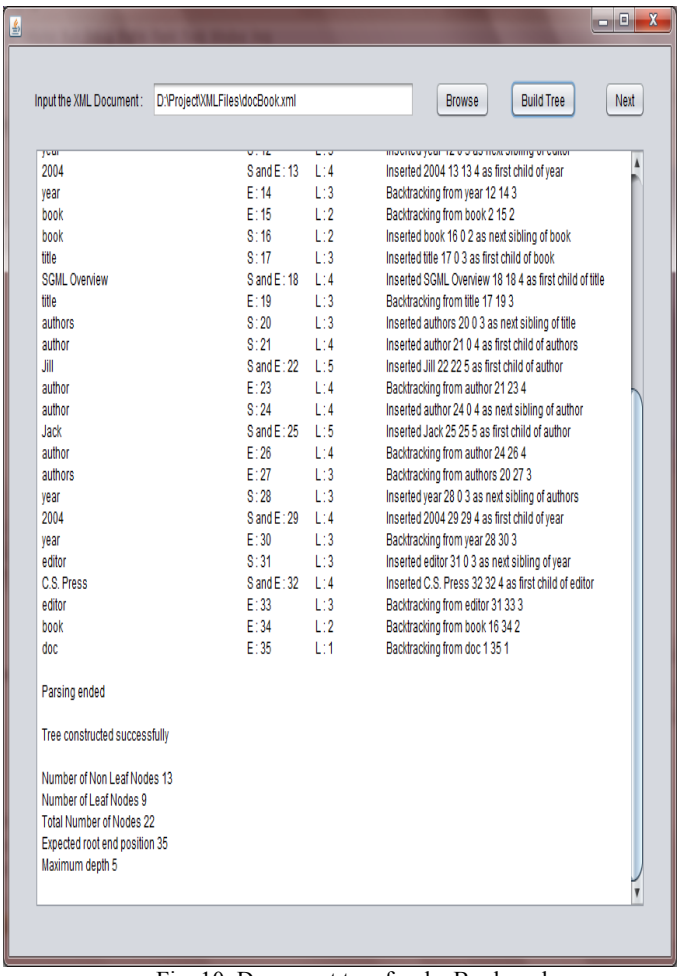


Fig. 10. Document tree for docBook.xml

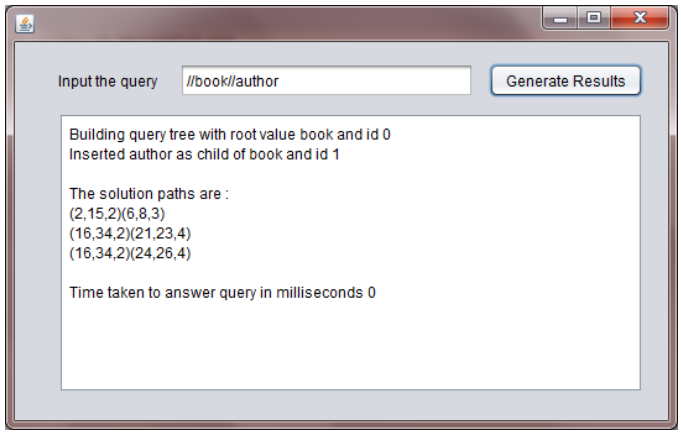


Fig. 11. Twig Results of //book//author

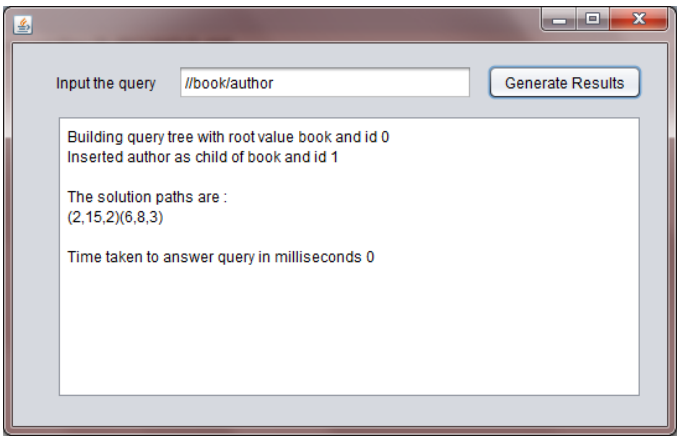


Fig. 12. Twig Results of //book/author

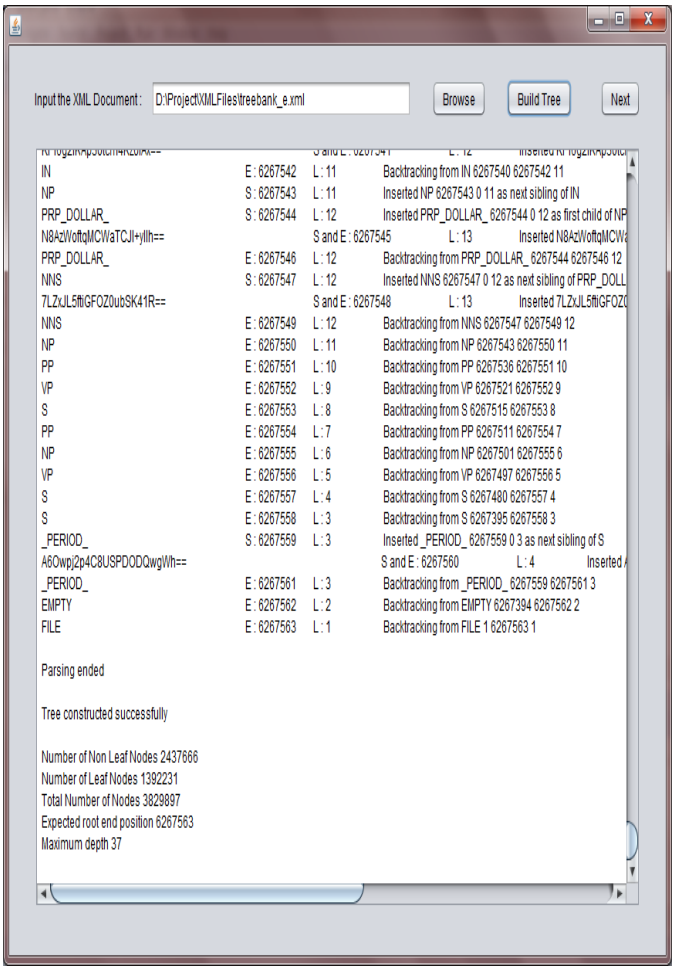


Fig. 13. Document tree for treebank_e.xml

ISBN: 978-1-61804-236-1

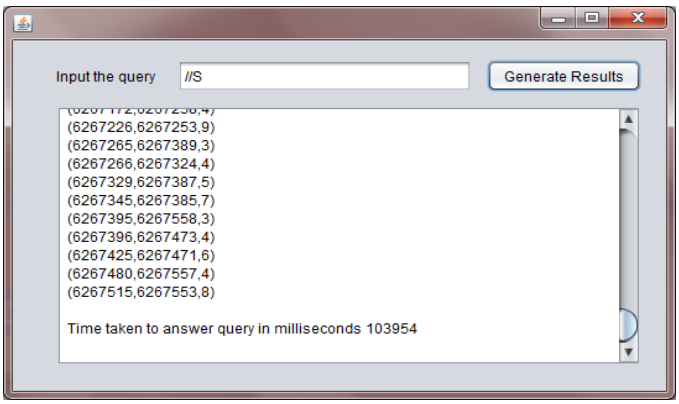


Fig. 14. Twig Results of //S

VIII. CONCLUSION

Holistic twig joins are critical operations for XML queries. In this paper, a novel approach for holistic computing of twig patterns using an algorithm called TwigMerge, which gracefully extends the I/O and CPU optimality to twigs with AD as well as PC edges, was presented. Analytical study was performed with regard to the validity and performance of the approach and its accompanying algorithms, and concluded with optimal I/O and optimal CPU on twigs with arbitrary AD and/or PC edges. This work supports only plain twigs which have only query nodes. As future work, the approach can be extended to boolean twigs or Btwigs i.e. twigs which support any arbitrary combination of AND/OR/NOT boolean predicates as well.

REFERENCES

[1] D. Che, T.W. Ling, W.C. Hou, "Holistic Boolean-Twig Pattern Matching for Efficient XML Query Processing", *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 11, pp. 2008-2024, Nov. 2012.

[2] N. Bruno, N. Koudas, and D. Srivastava, "Holistic Twig Joins: Optimal XML Pattern Matching", Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD' 02), pp. 310-321, June 2002.

[3] H. Jiang, H. Lu, W. Wang, "Efficient Processing of Twig Queries with OR-Predicates", Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD' 04), pp. 59-70, 2004.

[4] J. Lu, T. Chen, and T.W. Ling, "Efficient Processing of XML Twig Patterns with Parent Child Edges: A Look-ahead Approach", Proc. 13th ACM Int'l Conf. Information and Knowledge Management (CIKM' 04), pp. 533-542, Nov. 2004.

[5] T. Yu, T.W. Ling, and J. Lu, "twigstacklist¬: A Holistic Twig Join Algorithm for Twig Query with Not-Predicates on XML Data", Proc. 11th Int'l Conf. Database Systems for Advanced Applications (DASFAA' 06), pp. 249-263, 2006.

[6] Univ. of Washington XML Repository, http: //www.cs.washington.edu / research/xmldata sets/, 2012.

Divya Rajagopal completed her Bachelor's degree in Computer Science and Engineering from Amrita School of Engineering, Amrita Vishwa Vidyapeetham University, Coimbatore, Tamilnadu, India. She completed a 4-month internship with Infosys Ltd. and worked for a year as Systems Engineer in the Education and Research (E&R) department of Infosys Ltd., Mysore, India. She is currently pursuing her Master's degree in Computer Science and Engineering from Government College of Technology, Coimbatore, Tamilnadu, India.

Dr. Miraclin Joyce Pamila J.C. is an Assistant Professor (Senior Grade) in the Department of Computer Science and Engineering, Government College of Technology, Coimbatore, Tamilnadu, India. She received her Master and Doctoral degree in Computer Science and Engineering from Anna University, Chennai, India. Her fields of interest include Mobile Computing, Data Management Systems, Network Security and Recovery system design for mobile networks. She is a life member of ISTE. She teaches and guides courses at both B.E. and M.E. levels in Computer Science and Information Technology. She has published 30 technical papers in national and international conferences and journals.

Augmentation security of Cloud Computing via sequence unscented Kalman filtering

Mehdi Darbandi

Abstract: This paper can be organized into these subsections. In the first section, authors provide enough explanation about Cloud Computing concept, they discuss about advantages and disadvantages of this large-scale network. In the next section, authors mention practical uses of this technology, and for better explanation and easier realization they discuss about uses of this technology in IBM products and also discuss their decisions about what they want to do on this platform and by these resources on their future products. At the last part of this paper, authors purpose new kind of filter which named as "sequence unscented Kalman Filtering". Authors asserted that if we apply this evolutionary algorithm on cloud computing gateways and infrastructures we can estimate and predict the presence of hackers and spywares and Trojans and also we can estimate and predict the amount of power and resource consumptions, so that we can better satisfy user demands.

Keywords: Cloud Computing, Internet technology, Kalman filter.

Introduction: Most of secure applications today were usually developed first with their basic functionality, and security was added later, if at all, as an add–on extension or as additional, optional feature. If some already developed and operational application is to be enhanced with security, then the usual approach today is to invoke application programming interfaces (APIs) of some crypto library or use some, so called, crypto services provider. However, security tools and libraries today are not broadly available, sometimes not fully functional, and usually very complicated to use. Furthermore, in this process security functions are usually applied only to resources and functions of a specific application [10]. In addition, if an application offers some security services, then end-user has to configure various options and parameters prior to use of these security services. The procedures for that are usually inconvenient, especially for non-technical users. Finally, those applications are protected by additional external modules, like firewalls and virus scanners, after their installation and deployment.

Such add–on security extensions of applications, analysis of consequences and damages after network penetrations, recovery after destruction of resources, analysis of vulnerabilities of software modules for infection and other "post–factum" methods so far have shown their weaknesses and inadequate protection effects [10].

Security Provider provides security services to various components. The Provider is designed using the concept of generic security objects. Each generic object encapsulates security functions and attributes of some security service. The Provider transparently handles security credentials and hardware tokens, which are easy to integrate with other components.

Security protocols component comprises various network security protocols that provide authentication, authorization, secure communication, and identity verification services. These protocols are based on generic security objects, security standards and well-

established security technologies. Some of protocols are FIPS 196 strong authentication, Single-Sign-On, SAML authorization, and secure sessions protocol. These protocols also use Security Provider for various software-based or FIPS 201 (PIV) smart cards-based cryptographic functions [10].

Generic Security Server is another complex object which provides core components for implementation of Secure Application Servers supporting standard and extended security functions. All security functions are based on well-established security standards, technologies and protocols. Furthermore, several components, actions and libraries are available in this template in the form of Eclipse plug-ins in order to provide easy management of Secure Application Server, extendable with customized security functions, and several implemented actions for administration [10].

Impacts of Cloud Computing in IBM products:

IBM has multiple Cloud solutions to different problems, starting from simple SaaS applications such as CRM systems to complicated DB servers with different security tools. One of its Clouds called "SmartCloud" was developed for education purposes. SmartCloud provides services to design educational systems for schools and higher education without devoted staff or infrastructure. The IBM SmartCloud consists of a set of Cloud services for educators to follow and analyze student performance. In addition, it offers more effective research tools by maximizing resource availability; thus, it overcomes resource limitation in the local institutions' infrastructure. IBM SmartCloud provides the following solutions [10]:

- *SPSS Decision Management for Education: It is a Cloud based solution to analyze student information with different tools to identify students who will be enrolled in their institution by helping them maintain and succeed in their educational life and giving them appropriate information toward the right findings.*
- *Virtual Computing Lab (VCL): It provides a different services and tools via the Cloud for students and staff research. VCL (Averitt et al., 2007; Vouk, 2008; Mason Virtual Computing Lab, 2011) is simple to implement and maintain compared to other available solutions, flexibility, cost effective solution, and wide resources. VCL offers all of the three Cloud services: IaaS, PaaS, and SaaS. On the top is the infrastructure service which prevents students and staff from setting up any software or hardware on their computer while doing their assignment or research. VCL provides the following services for infrastructure [10]:*
  - *Compute resources, such as physical machines, virtual machines, and OS in the virtualization layer.*
  - *Network*
  - *Storage*
  - *Cloud Academy: This service provides the needed support to customers who wish to move to a Cloud and share their knowledge. It supports technical and business projects by allowing access to resources with the funding possibilities.*

IBM assists the institutions and companies in their process of moving to the Cloud by helping to build strategies, developing architecture, selecting the right workloads, determining a suitable deployment model—whether private, public, or hybrid Cloud—and managing their Cloud (Boss et al., 2007) [10].

ISBN: 978-1-61804-236-1

Sequence Unscented Kalman Filtering Algorithm:

Since Unscented Kalman Filter (UKF) is proposed by Julier and Uhlman, it has absolved many researchers to study it, and many kinds of new algorithms [2-6] with different accuracies have come out. Unlike the Extend Kalman Filter (EKF), which is based on the linearizing the nonlinear function by using Taylor series expansions, UKF uses the true nonlinear models and approximates a distribution of the state random variable [11]. Furthermore, it only needs a minimal set carefully chosen sample points, by which the posterior mean and covariance can be accurate to the second order for any nonlinearity, avoiding Jacobian's computation. If the priori random variable is Gaussian, the estimation of the posterior mean and covariance can could be accurate to the third order. It can be seen that in all the UKF algorithms it needs inversing the matrix in measurement update. The dimension of the inversed matrix is equal to that of measurement vector. If the dimension of the measurement is very large, so it could cost a great computing time. In order to decrease the computing cost and not to inverse the matrix, a sequence method is used to solve this problem [11].

In this paper, the sequence UKF is proposed. It deals with nonlinear stochastic system with linear measurement. Based on RBUKF and traditional Kalman Filter (KF), it deduced the special algorithm for the sequence UKF in case of the covariance matrix of measurement noise is diagonal matrix or not. In theory it is proven that the sequence UKF has the same estimation accuracy with RBUKF, but has lower computational cost. Simulation results verify the high performance of sequence UKF.

UKF MECHANISM:

UKF is used to solve the estimation problem for any nonlinear system. The considered nonlinear system is represented by [11]:

$$\begin{cases}\mathbf{x}_{k+1}=\mathbf{f}(\mathbf{x}_k)+\mathbf{v}_k\\ \mathbf{z}_k=\mathbf{h}(\mathbf{x}_k)+\mathbf{w}_k\end{cases}.$$

Where $\mathbf{x}_k$ and $\mathbf{z}_k$ denote the state vector with n-dimension and the measurement vector with m-dimension at step k, respectively. The nonlinear mapping $\mathrm{f}(\cdot)$ and $\mathrm{h}(\cdot)$ are assumed to be continuously differentiable with respect to $\mathbf{x}_k$. Moreover, $\mathbf{v}_k \sim \mathrm{N}(\mathbf{v}_k;\mathbf{0},\mathbf{Q}_k)$ denote the process noise with n-dimension. $\mathbf{w}_k \sim \mathrm{N}(\mathbf{w}_k;\mathbf{0},\mathbf{R}_k)$ denote the measurement noise with m-dimension. $\mathbf{v}_k$ and $\mathbf{w}_k$ are independent of each other.

Like Kalman Filter (KF), UKF is also a minimum mean-square error estimator (MMSE). For system (1), the mechanism of MMSE is time-update and measurement-update as follows [11]:

Time-update:

$$\mathbf{x}_{k/k-1}=E[\mathrm{f}(\mathbf{x}_{k-1})]$$

$$\mathbf{P}_{k/k-1}=E[\mathbf{e}_k\mathbf{e}_k^T]$$

Measurement-update:

$$\hat{\mathbf{x}}_k=\mathbf{x}_{k/k-1}+\mathbf{W}_k\boldsymbol{\upsilon}_n$$

$$\boldsymbol{\upsilon}_k=\mathbf{z}_k-\hat{\mathbf{z}}_k$$

$$\hat{\mathbf{z}}_k=E[\mathrm{h}(\mathbf{x}_k)]$$

$$\mathbf{P}_k=\mathbf{P}_{k/k-1}-\mathbf{W}_k\mathbf{S}_k\mathbf{W}_k^T$$

Where $\mathbf{e}_k=\mathbf{x}_k-\mathbf{x}_{k/k-1}$, the weight matrix $\mathbf{W}_k$ is chosen to minimize the trace of the updated covariance $\mathbf{P}_k$. Its value is calculated from [11]:

$$\mathbf{W}_k=\mathbf{P}_k^{xz}\mathbf{S}_k^{-1}$$

Where $\mathbf{P}_k^{xz}$ is the cross covariance between $\mathbf{e}_k$ and $\boldsymbol{\upsilon}_k$, $\mathbf{S}_k$ is the covariance of $\boldsymbol{\upsilon}_k$.

UKF is based on the mechanism above. By applying the unscented transformation (UT) to

ISBN: 978-1-61804-236-1

a number of chosen sigma points, $\mathbf{x}_{k/k-1}$, $\mathbf{P}_{k/k-1}$, $\mathbf{S}_k$ and $\mathbf{P}_k^{xz}$ can be approximately expressed by the linear composition of the transformed sigma points. So UKF solves the nonlinear estimation problem using MMSE mechanism. When measurement equation in system is linear, it changes to system as follows [11]:

$$\begin{cases}\mathbf{x}_{k+1} = \mathrm{f}(\mathbf{x}_k) + \mathbf{v}_k \\ \mathbf{z}_k = \mathbf{H}_k x_k + \mathbf{w}_k\end{cases}$$

To solve the estimation problem for system, it only needs transformed sigma points to estimate $\mathbf{x}_{k/k-1}$ and $\mathbf{P}_{k/k-1}$, and $\mathbf{S}_k$ and $\mathbf{P}_k^{xz}$ can be computed accurately. So the UKF algorithm can be reduced to RBUKF algorithm. Compared to UKF, RBUKF is not less computational cost, but also higher accuracies. The difference part between the UKF and RBUKF is measurement-update. For comparison, it only gives the measurement-update of RBUKF as follows [11]:

$$\mathbf{P}_{zz,k} = \mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}_k^T + \mathbf{R}_k$$

$$\mathbf{P}_{k,xz} = \mathbf{P}_{k|k-1} \mathbf{H}_k^T$$

$$\mathbf{K}_k = \mathbf{P}_{xz,k|k-1} \mathbf{P}_{zz,k}^{-1}$$

$$\hat{\mathbf{x}}_k = \mathbf{x}_{k|k-1} + \mathbf{K}_k (\mathbf{z}_k - \mathbf{z}_{k|k-1})$$

$$\mathbf{P}_k = \mathbf{P}_{k|k-1} - \mathbf{K}_k \mathbf{P}_{zz,k} \mathbf{K}_k^T$$

SEQUENCE UKF ALGORITHM

No matter UKF or RBUKF, it needs to inverse the matrix in measurement-update. If the dimension of the measurement vector is very large, it could be a great computational cost for compute gain matrix $\mathbf{K}_k$. In order to avoid inversing the matrix in computing $\mathbf{K}_k$, the sequence UKF will deal with ever component of measurement vector one by one instead of the vector at one time. This method needs not to inverse the matrix and can greatly decrease computational cost. For system, it deduced the sequence UKF as follows [11].

Theory I: For system, the measurement-update in UKF can be computed as follows:

$$\mathbf{K}_k^i = \mathbf{P}_k^{i-1} \mathbf{H}_k^{iT} (\mathbf{H}_k^i \mathbf{P}_k^{i-1} \mathbf{H}_k^{iT} + R_k^i)^{-1}$$

$$\mathbf{x}_k^i = \mathbf{x}_k^{i-1} + \mathbf{K}_k^i (z_k^i - \mathbf{H}_k^i \mathbf{x}_k^{i-1})$$

$$\mathbf{P}_k^i = \mathbf{P}_k^{i-1} - \mathbf{K}_k^i \mathbf{H}_k^i \mathbf{P}_k^{i-1} \quad i = 1, 2, \cdots, m$$

Where $\mathbf{H}_k^i$ is the i-th row in $\mathbf{H}_k$, z_k^i is the i-th scalar in the measurement vector at step k, R_k^i is the i-th element in diagonal of $\mathbf{R}_k$ [11].

Proof:

Rewrite the measurement equation in system, it gets [11]:

$$\begin{bmatrix} z_k^1 \\ z_k^1 \\ \vdots \\ z_k^m \end{bmatrix} = \begin{bmatrix} \mathbf{H}_k^1 \\ \mathbf{H}_k^2 \\ \vdots \\ \mathbf{H}_k^m \end{bmatrix} \mathbf{x}_k + \begin{bmatrix} v_k^1 \\ v_k^2 \\ \vdots \\ v_k^m \end{bmatrix}$$

Because of the form of, the measurement vector $\mathbf{z}_k$ can be seen as z_k^i ($i = 1, 2, \cdots, m$) one by one to update the time-update equations. It must be noticed that when z_k^i one by one instead of $\mathbf{z}_k$ updates the equations in measurement-update, it implies that the state equation is invariant for every z_k^i at step k. So the measurement-update in UKF at step k is equivalent to the filtering problem of new linear systems as follows [11]:

$$\begin{cases}\mathbf{x}_k^i = \mathbf{x}_k^{i-1} \\ z_k^i = \mathbf{H}_k^i \mathbf{x}_k^i + w_k^i \quad (i = 1, 2, \cdots, m)\end{cases}$$

Where $w_k^i \sim \mathrm{N}(w_k^i; 0, R_k^i)$, is the equivalent measurement noise. The initial value of the filter is $\mathbf{x}_k^0 = \mathbf{x}_{k/k-1}$ and $\mathbf{P}_k^0 = \mathbf{P}_{k/k-1}$.

So the equivalent linear filter can be easily derived from classic Kalman Filter equations as follows:

$$\mathbf{x}_k^{i/i-1} = \mathbf{x}_k^{i-1}$$

$$\mathbf{x}_k^i = \mathbf{x}_k^{i/i-1} + \mathbf{K}_k^i (z_k^i - \mathbf{H}_k^i \mathbf{x}_k^{i/i-1})$$

$$\mathbf{K}_k^i = \mathbf{P}_k^{i/i-1} \mathbf{H}_k^{iT} (\mathbf{H}_k^i \mathbf{P}_k^{i/i-1} \mathbf{H}_k^{iT} + R_k^i)^{-1}$$

$$\mathbf{P}_k^{i/i-1} = \mathbf{I}\mathbf{P}_k^{i-1}\mathbf{I}^T$$

$$\mathbf{P}_k^i = (\mathbf{I} - \mathbf{K}_k^i \mathbf{H}_k^i)\mathbf{P}_k^{i/i-1}$$

Or $(\mathbf{P}_k^i)^{-1} = (\mathbf{P}_k^{i/i-1})^{-1} + \mathbf{H}_k^{iT}(R_k^i)^{-1}\mathbf{H}_k^i$

Substituting these recent equations results in the following, and substituting previous equations in each other, results in next equations. So it obtains the new equations for the measurement-update in UKF as follows:

$$\mathbf{x}_k^i = \mathbf{x}_k^{i-1} + \mathbf{K}_k^i(z_k^i - \mathbf{H}_k^i \mathbf{x}_k^{i-1})$$

$$\mathbf{K}_k^i = \mathbf{P}_k^{i-1}\mathbf{H}_k^{iT}(\mathbf{H}_k^i \mathbf{P}_k^{i-1}\mathbf{H}_k^{iT} + R_k^i)^{-1}$$

$$\mathbf{P}_k^i = (\mathbf{I} - \mathbf{K}_k^i \mathbf{H}_k^i)\mathbf{P}_k^{i-1}$$

Or $(\mathbf{P}_k^i)^{-1} = (\mathbf{P}_k^{i-1})^{-1} + \mathbf{H}_k^{iT}(R_k^i)^{-1}\mathbf{H}_k^i$

By theory I and the UKF mechanism, it obtains the sequence UKF algorithm as follows:

Calculate sigma points [11]:

$$\begin{cases} \boldsymbol{\chi}_{i,k-1} = \mathbf{x}_{i,k-1} & i = 0, \\ \boldsymbol{\chi}_{i,k-1} = \mathbf{x}_{i,k-1} + (\sqrt{(L+\lambda)\mathbf{P}_{k-1}})_i & i = 1,\cdots,L, \\ \boldsymbol{\chi}_{i,k-1} = \mathbf{x}_{i,k-1} - (\sqrt{(L+\lambda)\mathbf{P}_{k-1}})_{i-L} & i = L+1,\cdots,2L, \end{cases}$$

Time-update:

$$\boldsymbol{\chi}_{k/k-1} = \mathrm{f}(\boldsymbol{\chi}_{k-1})$$

$$\mathbf{x}_{k/k-1} = \sum_{i=0}^{2n} W_i^{(m)} \boldsymbol{\chi}_{i,k/k-1}$$

$$\mathbf{P}_{k/k-1} = \sum_{i=0}^{2n} W_i^{(c)} (\boldsymbol{\chi}_{i,k/k-1} - \mathbf{x}_{k|k-1})(\boldsymbol{\chi}_{i,k/k-1} - \mathbf{x}_{k/k-1})^T + \mathbf{Q}_k$$

$$z_{k/k-1}^i = \mathbf{H}_k^i \mathbf{x}_{k/k-1}^i \quad (i = 1,2,\cdots,m)$$

Measurement-update:

$$\mathbf{P}_k^0 = \mathbf{P}_{k/k-1}$$

$$\mathbf{x}_k^0 = \mathbf{x}_{k/k-1}$$

$$\mathbf{x}_k^i = \mathbf{x}_k^{i-1} + \mathbf{K}_k^i(z_k^i - \mathbf{H}_k^i \mathbf{x}_k^{i-1})$$

$$\mathbf{K}_k^i = \mathbf{P}_k^{i-1}\mathbf{H}_k^{iT}(\mathbf{H}_k^i \mathbf{P}_k^{i-1}\mathbf{H}_k^{iT} + R_k^i)^{-1}$$

$$\mathbf{P}_k^i = (\mathbf{I} - \mathbf{K}_k^i \mathbf{H}_k^i)\mathbf{P}_k^{i-1} \quad (i = 1,2,\cdots,m)$$

$$\mathbf{x}_k = \mathbf{x}_k^m$$

$$\mathbf{P}_k = \mathbf{P}_k^m$$

Where $W_0^{(c)} = \lambda/(L+\lambda) + (1-\alpha^2+\beta)$,

$W_i^{(m)} = W_i^{(c)} = 1/[2(L+\lambda)]$, $i = 1,2,\cdots,2L$.

ALGORITHM PERFORMANCE ANALYSIS:

Filtering accuracy Analysis:

For system, in order to comparison of RBUKF and the sequence UKF, it needs to get the filtering covariance respectively. Firstly, it is to derived covariance $\mathbf{P}_k$ of RBUKF. By definition, it calculates covariance $\mathbf{P}_k$ of RBUKF as follows:

$$\mathbf{P}_k = E[(\mathbf{x}_k - \hat{\mathbf{x}}_k)(\mathbf{x}_k - \hat{\mathbf{x}}_k)^T]$$

Substitute this equation in previous equations results in [11]:

$$\mathbf{P}_k = \mathbf{P}_{k/k-1} - \mathbf{P}_{k/k-1}\mathbf{H}_k^T\mathbf{K}_k^T - \mathbf{K}_k\mathbf{H}_k\mathbf{P}_{k/k-1} + \mathbf{K}_k\mathbf{H}_k\mathbf{P}_{k/k-1}\mathbf{H}_k^T\mathbf{K}_k^T$$

By previous equations, it obtains

$$\mathbf{K}_k = \mathbf{P}_{xz,k/k-1}\mathbf{P}_{zz,k/k-1}^{-1} = \mathbf{P}_{k/k-1}(\mathbf{H}_k\mathbf{P}_{k/k-1}\mathbf{H}_k^T + \mathbf{R}_k)^{-1}$$

Substituting these recent equations results in:

$$\mathbf{P}_k = \mathbf{P}_{k/k-1} - \mathbf{P}_{k/k-1}\mathbf{H}_k^T(\mathbf{H}_k\mathbf{P}_{k/k-1}\mathbf{H}_k^T + \mathbf{R}_k)^{-1}\mathbf{H}_k\mathbf{P}_{k/k-1} = (\mathbf{P}_{k/k-1}^{-1} + \mathbf{H}_k^T\mathbf{R}_k^{-1}\mathbf{H}_k)^{-1}$$

By inversing both sides of this equation, it obtains [11]:

$$\mathbf{P}_k^{-1} = \mathbf{P}_{k-1/k}^{-1} + \mathbf{H}_k^T\mathbf{R}_k^{-1}\mathbf{H}_k$$

Because of:

$$\mathbf{H}_k = [\mathbf{H}_k^{1T}\ \mathbf{H}_k^{2T}\ \cdots\ \mathbf{H}_k^{mT}]^T$$

$$\mathbf{R}_k^{-1} = (diag(R_k^1\ R_k^2\ \cdots\ R_k^m))^{-1} = diag((R_k^1)^{-1}\ (R_k^2)^{-1}\ \cdots\ (R_k^m)^{-1})$$

Substituting these recent equations in each other, and doing matrices multiplication, it gets

$$\mathbf{P}_k^{-1} = \mathbf{P}_{k-1/k}^{-1} + \mathbf{H}_k^T\mathbf{R}_k^{-1}\mathbf{H}_k = \mathbf{P}_{k-1/k}^{-1} + \sum_{i=1}^{m}\mathbf{H}_k^{iT}(R_k^i)^{-1}\mathbf{H}_k^i$$

Secondly, it calculates the covariance of the sequence UKF. By the measurement-update equation, it easily gets

$$
\begin{aligned}
(\mathbf{P}_k^m)^{-1} &= (\mathbf{P}_k^{m-1})^{-1} + \mathbf{H}_k^{mT}(R_k^m)^{-1}\mathbf{H}_k^m \\
&= (\mathbf{P}_k^{m-2})^{-1} + (\mathbf{H}_k^{m-1})^T(R_k^{m-1})^{-1}\mathbf{H}_k^{m-1} \\
&+ \mathbf{H}_k^{mT}(R_k^m)^{-1}\mathbf{H}_k^m \\
&\cdots \\
&= (\mathbf{P}_k^0)^{-1} + \sum_{i=1}^{m}(\mathbf{H}_k^{i-1})^T(R_k^{i-1})^{-1}\mathbf{H}_k^{i-1}
\end{aligned}
$$

Substituting previous equations in this equation results in the covariance of the sequence UKF as follows:

$$
(\hat{\mathbf{P}}_k)^{-1} = = (\mathbf{P}_{k/k-1})^{-1} + \sum_{i=1}^{m}(\mathbf{H}_k^i)^T(R_k^i)^{-1}\mathbf{H}_k^i
$$

Compare these two recent equations, it can be seen that the covariance of the sequence UKF is equal to that of RUKF, which means that the accuracies of the two filters are the same in theory [11].

Computational Complexity Analysis:

In the sequence UKF algorithm, because $\mathbf{H}_k^i\mathbf{P}_k^{i-1}\mathbf{H}_k^{iT} + R_k^i$ is a scalar, so this algorithm has successfully converted the inversion of m-dimension matrix into m times division. The computational complexity has been greatly decreased. In order to compare the computational complexity between the sequence UKF and RBUKF, for their time-update algorithms are the same, here only gives the comparison results of their measurement-update algorithms in table I. From this table, it can be seen that number of calculating times in RBUKF algorithm contains the third order of measurement dimension, while the sequence UKF has only second order component. So when the measurement dimension is large, the computational cost in the sequence UKF will be much less [11].

TABLE I
CALCULATION TIMES COMPARISON

Algorithm	*Num of* $\times\div$	*Num of* $+-$
The sequence UKF	$(5m-1)n^2 + 2m^2n + \frac{2}{3}m^3 + m^2 + \frac{4}{3}m$	$5mn^2 + (m^2-4m)n + \frac{m^3}{2} + \frac{m}{2}$
RBUKF	$4mn^2 + 4mn + m$	$4mn^2$

NUMERICAL SIMULATION:

In order to show the efficiency of the sequence UKF, it is applied to an example system in comparison with the RBUKF. Estimation performance and computational complexity of the filters are evaluated with Monte Carlo simulations [11].

The numerical example considered in this section is a fifth-order nonlinear model given by system, with four-dimension measurement.

$$
\mathbf{x}_{k+1} = \begin{bmatrix} \sin x_{1,k}\cos x_{2,k} + 0.5x_{2,k} - 0.1(x_{3,k})^2 \\ \sin x_{1,k} + (\sin x_{2,k})^2 - 0.1x_{5,k} \\ \cos x_{1,k} + \exp(-x_{2,k}) + 0.1(x_{3,k})^2 \\ \sin x_{5,k} + \cos^2(x_{4,k}) - 0.5x_{2,k} \\ \sin x_{4,k} + \cos^2(x_{3,k}) - 0.1x_{1,k} \end{bmatrix} + \begin{bmatrix} v_{1,k} \\ v_{2,k} \\ v_{3,k} \\ v_{4,k} \\ v_{5,k} \end{bmatrix}
$$

$$
\mathbf{H}_k = \begin{bmatrix} 0.1 & -0.2 & 0 & 0 & 0 \\ 0.15 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & -0.5 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}
$$

The covariance matrices of $\mathbf{v}_k$ and $\mathbf{w}_k$ are:
$\mathbf{Q}_k = 0.0001\mathbf{I}_5$, $\mathbf{R}_k = 0.01\mathbf{I}_4$.

The initial conditions for the system and the filters are [11]:

$$x_{1,0} = x_{2,0} = x_{3,0} = x_{4,0} = x_{5,0} = 0.5$$

$$\hat{x}_{1,0} = \hat{x}_{2,0} = \hat{x}_{3,0} = \hat{x}_{4,0} = \hat{x}_{5,0} = 2$$

And the initial covariance matrix is chosen as $\hat{\mathbf{P}}_0 = 100^2\mathbf{I}_5$.

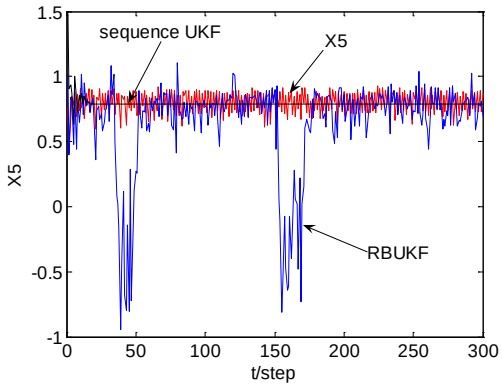


Fig.1 Comparison of state x_5

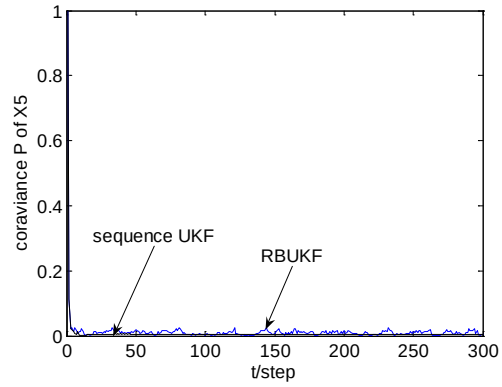


Fig.2 Comparison covariance of x_5

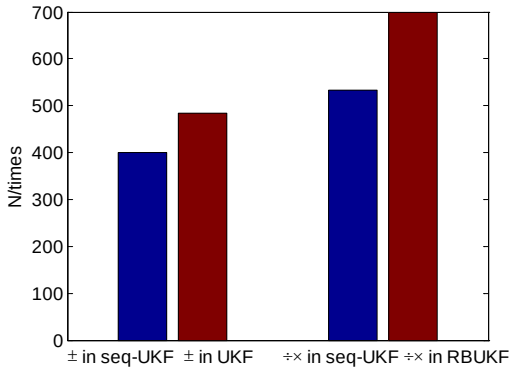


Fig.3 comparison of computational complexity

For briefness, it only shows the filtering results of state x_5. From fig. 1, it can be seen the filtering accuracy of the sequence UKF is same with that of RBUKF, but curve of the RBUKF is more fluctuant than that of the sequence UKF, which means that the sequence UKF maybe has a better filtering performance in actually usage. Fig. 2 is the filtering covariance of the two filters. It can be seen that the filtering covariance is almost equal to each other. Fig. 3 is the comparison of the computational complexity between the two filters for the example. It shows that the number of multiplication and division in the sequence UKF is much smaller than that in RBUKF, so is it for the number of addition and subtraction. It verifies that the computational complexity of sequence is less than RBUKF. For the computational cost of RBUKF is less than traditional UKF, So the sequence UKF will have great advantage in comparison with UKF [11].

Conclusion:

This paper mainly about adapting Sequence Unscented Kalman Filtering algorithm into cloud computing technology. For satisfying reader about this adaption, authors in the first section discuss about cloud computing and all different aspects of this technology. They define and explain most of the important parameters that are deal with cloud computing. In the next section of this paper, authors provide some information about comprehensive uses of cloud computing, especially they discuss more about uses of cloud computing in IBM products and discuss about future works of this company on the basis of cloud computing. To this end readers should acquire general realization about cloud computing and its applications. In the third section of this paper, authors reveal new filter and proof it by mathematical relations. They claim that, if this algorithm were implemented on cloud platforms and infrastructures, users face with excellent defense and have very good privacy over their information and documents. Also by the means of this innovative filter, we can

ISBN: 978-1-61804-236-1

estimate and predict about existence of hackers or crackers in such large-scale network.

References:

[1] Mehdi Darbandi "Applying Kalman Filtering in solving SSM estimation problem by the means of EM algorithm with considering a practical example"; published by the Journal of Computing – **Springer**, 2012; USA.

[2] Mehdi Darbandi; "Comparison between miscellaneous platforms that present for cloud computing and accreting the security of these platforms by new filter"; published by the Journal of Computing – **Springer,** 2012; USA.

[3] Mehdi Darbandi; "New and novel technique in designing electromagnetic filter for eliminating EMI radiations and optimization performances"; published by the Journal of Computing - **Springer**, 2012; USA.

[4] Mehdi Darbandi; "Developing concept of electromagnetic filter design by considering new parameters and use of mathematical analysis"; published by the Journal of Computing - **Springer**, 2012; USA.

[5] Mehdi Darbandi; "Is the cloud computing real or hype Affirmation momentous traits of this technology by proffering maiden scenarios"; published by the Journal of Computing – **Springer**, 2012; USA.

[6] Mehdi Darbandi; "Measurement and collation overriding traits of computer networks and ascertainment consequential exclusivities of cloud computing by the means of Bucy filtering"; published by the Journal of Computing - **Springer**, 2012; USA.

[7] Mehdi Darbandi; "Unabridged collation about multifarious computing methods and outreaching cloud computing based on innovative procedure"; published by the Journal of Computing - **Springer**, 2012; USA.

[8] Mehdi Darbandi; "Scrutiny about all security standards in cloud computing and present new novel standard for security of such networks"; published by the Journal of Computing - **Springer**, 2012; USA.

[9] Microsoft's Accessible Technology Vision and Strategy; September 2011.

[10] PhD. Thesis of Abdul Ghafoor Abbasi*; School of Information and Communication Technologies (ICT), KTH; Stockholm, Sweden; 2011.*

[11] Hui-ping Li, De-min Xu and Fu-bin Zhang ; "Sequence Unscented Kalman Filtering Algorithm".

[12] Mehdi Darbandi; "Appraising the role of cloud computing in daily life and presenting new solutions for stabilization of this technology"; published by the Journal of Computing - **Springer**, 2012; USA.

[13] Mehdi Darbandi; "Cloud Computing make a revolution in economy and Information Technology"; published by the Journal of Computing - **Springer**, 2012; USA.

[14] Mehdi Darbandi; "Considering the high impact of gettering of silicon on fabrication of wafer designing and optimize the designing with new innovative solutions"; published by the Journal of Computing – **Springer**, 2012; USA.

The Potential Role of Case-based Reasoning in Myocardial SPECT Perfusion

Shymaa H. ElRefaie and Abdel-Badeeh M. Salem

Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
Shymaa.elrefaie@hotmail.com abmsalem@yahoo.com absalem@cis.asu.edu.eg

***Abstract*—**myocardial perfusion SPECT imaging is an established noninvasive method for the functional assessment of coronary artery disease. Visual interpretation of perfusion scintigrams requires experienced readers and is associated with interobserver variability. Thus, computer based methods have been developed that support decisions with quantitative image analysis. Case based reasoning (CBR) approach presents a foundation for a robust methodology of building intelligent computer-aided diagnoses systems in medical domain. CBR system searches a library of patient cases to find the ones that best match those of the patient study being analyzed. The common findings from these cases, such as coronary angiography results, are then used to assist the diagnostician's interpretation. This paper presents the potential role of the CBR methodology in diagnosis of myocardial SPECT perfusion .In addition, the paper discusses the technical challenges and difficulties in developing the CBR based systems from the knowledge engineering point of view.

***Keywords*—**Artificial intelligence, Case Based Reasoning, Medical informatics, Myocardial Perfusion, Nuclear medicine; SPECT.

I. INTRODUCTION

Reasoning from experience is a natural way of human thinking, one remembers an apparently similar situation, what one has done and what the outcome has been; accordingly one acts in the present situation. CBR draws from this paradigm and tries to formalize it for use on the computer. CBR means reasoning from experiences or "old cases" in an effort to solve problems, critique solutions, and explain anomalous situations .CBR is the scientific method (or collection of methods) to imitate and enhance, if possible, this human behavior to find useful and applicable old cases and to reuse them either directly or after adaptation. In addition, the success of adaptation has to be verified and cases have to be collected for future use.CBR, as a computational intelligence method ,assumes a memory model for representing, indexing and organizing past cases and a process model for retrieving and modifying old cases and assimilating new ones [1,2].

1) There are two styles of CBR; problem solving style and interpretive style. Problem solving style can support a variety of tasks including planning, diagnosis and design (e.g. Medicine [2] and Industry [3]). The interpretive style is useful for (a) situation classification, (b) evaluation of solution, (c) argumentation, (d) justification of solution interpretation or plan and (e) the projection of effects of a decision of plan. Lawyers and managers making strategic decisions use the interpretive style [4, 5]. CBR has already been applied in a number of different applications in medicine. Some real CBR-systems are: CASEY that gives a diagnosis for the heart disorders [1], GS.52 which is a diagnostic support system for dysmorphic syndromes, NIMON is a renal function monitoring system, COSYL that gives a consultation for a liver transplanted patient [6] and ICONS that presents a suitable calculated antibiotics therapy advise for intensive care patients [7].

Single Photon Emission Computed Tomography (SPECT) scintigraphy is a noninvasive diagnostic method for the assessment of presence and severity of coronary artery disease (CAD) [8]. Artificial

intelligence (AI) techniques have been utilized to develop automated diagnostic systems with the aim to support human readers and improve the diagnostic accuracy [9]. These methods use the individual diagnostic information of both normal and abnormal images of a case library and thus can potentially offer better diagnostic accuracy. In this paper we focus our discussion around the potential role of the usage of CBR approach for diagnosis of myocardial SPECT perfusion

II. Case Based Reasoning (CBR) Methodology

CBR is an analogical reasoning method provides both a methodology for problem solving and a cognitive model of people. The "case" is a list of features that lead to a particular outcome (e.g. The information on a patient history and the associated diagnosis). CBR means reasoning from experiences or "old cases" in an effort to solve problems, critique solutions, and explain anomalous situations. It is consistent with much that psychologist have observed in the natural problem solving that people do. People tend to be comfortable using CBR methodology for decision making, in dynamically changing situations and other situations were much is unknown and when solutions are not clear.

From the knowledge computing point of view, CBR refers to a number of intelligent algorithms and techniques that can be used to record and index cases and then search them to identify the ones that might be useful in solving new cases when they are presented. In addition, there are techniques that can be used to modify earlier cases to better match new cases and other techniques to synthesize new cases when they are needed [1,2]. CBR system will search its "Case-Memory" for an existing case that matches the input problem specification. If we are lucky (our luck increases as we add new cases to the system), we will find a case that exactly matches the input problem and goes directly to a solution. If we are not lucky, we will retrieve a case that is similar to our input situation but not entirely appropriate to provide as a completed solution. The system must find and modify small portions of the retrieved case that do not meet the input specification. This process is called "case-adaptation". The result of case adaptation process is (a) completed solution, and (b) generates a new case that can be automatically added to the system's case-memory for future use. The CBR process can be represented by a schematic cycle, as shown in Figure 1 [10].

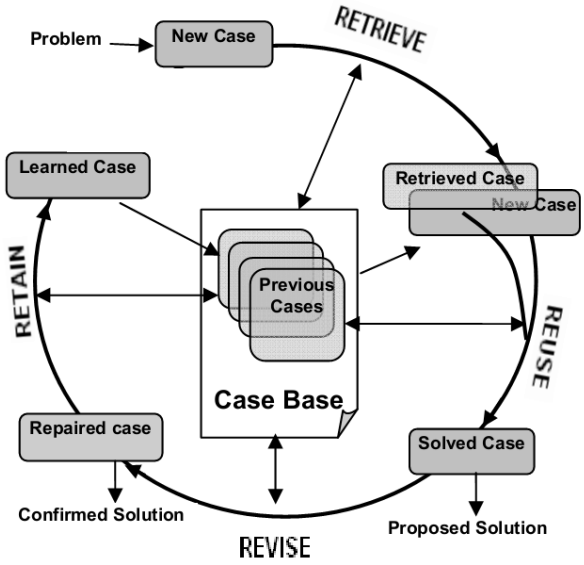


Figure 1. CBR cycle. The figure is introduced by Aamodtand Plaza [10]

The advantages of CBR in medical domain have been identified and explored in several research works i.e. in [11, 12, and 13]. Several motivation of applying CBR in medical domain can be identified as:

1. CBR method can work in a way close to human reasoning e.g. solves a new problem applying previous experiences.
2. Knowledge elicitation is another problem in some medical domain, as human behavior is not always predictable. Even for an experienced clinician might have difficulty to articulate their knowledge explicitly. Sometimes they make assumptions and predictions based on experiences or old cases. Using CBR this knowledge elicitation bottleneck can be overcome.
3. CBR can be used when there are no sets of rules or a model.
4. Sometimes it is possible to identify features for the success or failure of a case. This would help to reduce the repetition of mistakes in future.
5. The knowledge in medical domain is growing with time so it is important the system can learn new knowledge. Many of the AI systems failed to continue because of the lack of this type of maintenance. CBR system can learn by adding new cases into the case base.

ISBN: 978-1-61804-236-1

6. The cases in the case base can be used for the follow up of the treatment and also for training purposes of the less experience clinicians.

III. Technical Challenges and Difficulties in Developing CBR Systems

CBR approach addresses the problems found in the traditional AI techniques, e.g. the problems of knowledge acquisition, remembering, performance, robust and maintenance[1,6] But, from the implementation point of view, the knowledge engineers are facing many technical challenges and difficulties in developing the CBR systems in medical domain. This section discusses these issues.

A. Case Representation:

The case is a list of features that lead to a particular outcome (e.g. The information on a patient history and the associated diagnosis).Determining the appropriate case features is the main knowledge engineering task in CBR systems. The task involves; (a) defining the terminology of the domain and (b) gathering representative examples of problem solving by the expert. Representations of cases can be in any of several forms; predicate representations, frame representations and representations resembling database entries.

B. Case Indexing:

The CBR system derives its power from its ability to retrieve relevant cases quickly and accurately from its memory. Figuring out when a case should be selected for retrieval in similar future situations is the goal of the case indexing process. Building a structure or process that will return the most appropriate case (from the case memory) is the goal of the retrieval process. Case indexing process usually falls into one of three approaches: nearest neighbor, inductive and knowledge-guided.

C. Case Memory Organization and Retrieval:

Once cases are represented and indexed, they can be organized into an efficient structure for retrieval. Most case memory structures fall into a range between purely associative retrieval, where any or all of the features of a case are indexed independently of the other features and purely hierarchical retrieval, where case features are highly organized into a general-to-specific a concept structure. Nearest-neighbor matching techniques are considered associative because they have no real-memory organization. Discrimination nets are more of a cross between associative and hierarchical because they have some structure to the net but greater retrieval flexibility because they have a greater number of links between potential indexing features. Decision trees are an example of purely hierarchical memory organization. The type of memory organization is related to the amount of knowledge available to perform indexing and the retrieval needs of the system. If flexibility is required because one case library is being used for several retrieval tasks, a more associative approach is often used. When the retrieval task is well defined, a hierarchical approach is used because of the advantages in retrieval time the hierarchical approaches have over associative approaches.

D. Case Adaptation:

It is difficult to define a single generically applicable approach to perform case adaptation, because adaptation tends to be problem specific. Most existing CBR systems achieve case adaptation for the specific problem domains they address by encoding adaptation knowledge in the form of a set of adaptation rules or domain model. Adaptation rules are then applied to a retrieved case to transform it into a new case that meets all of the input problem's constraints. More recent applications have successfully used pieces of existing cases in memory to perform adaptations. In problem domains where it is difficult to codify enough rule-like knowledge to let board adaptation be done, using pieces of cases is the best, if not the only alternative. And, even if cases can't be adapted by the computer, at least the system has provided the human "adapter" with a significant starting point.

E. Learning and Generalization:

as cases accumulates, case generalization can be used to define prototypical cases that embody the major features of a group of specific cases, and those prototypical cases can be stored with the specific cases, improving the accuracy of the system in the long run. In addition, inductive-case analysis research

ISBN: 978-1-61804-236-1

is being done to build domain theories in areas where even the experts don't understand how the underlying processes in their domain.

IV. Single Photon Emission Computer Tomography (SPECT)

Myocardial perfusion imaging (MPI) is a form of functional cardiac imaging, used for the diagnosis of ischemic heart disease. The underlying principle is that under conditions of stress, diseased myocardium receives less blood flow than normal myocardium. MPI is one of several types of cardiac stress test [8].

SPECT is a nuclear medicine tomographic imaging technique using gamma rays. It is very similar to conventional nuclear medicine planar imaging using a gamma camera. However, it is able to provide three dimensional (3D) information. This information is typically presented as cross-sectional slices through the patient, but can be freely reformatted or manipulated as required (see Fig. 2).

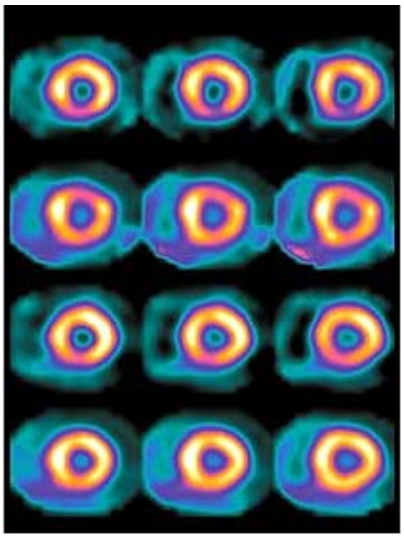


Figure 2. SPECT image of short-axis slices [14]

Cardiac stress test is a test used in medicine and cardiology, to measure the heart ability to external stress, when the measurement is in a controlled clinical environment. The stress response is actually induced by exercise or stimulated with drugs. Cardiac stress tests measure the coronary circulation at rest with that observed during maximum physical exertion. You detect any imbalances and stress of blood flow to the myocardium. The results are interpreted as a reflection on the general physical condition of the test patient.

For diagnostic of nuclear medicine, SPECT imaging performed after stress reveals the distribution of the radiopharmaceutical, and therefore the relative blood flow to the different regions of the myocardium [15]. Diagnosis is made by comparing stress images to a further set of images obtained at rest. Cardiologist will evaluate heart perfusion levels by comparing a patient under stress conditions, relative to normal resting perfusion levels, by inspecting the SPECT images. This task is tedious for cardiologists. Figure 3 shows a nuclear image of a heart with CAD at rest and after exercise. The change in color shows that less blood is reaching a part of the heart muscle [16].

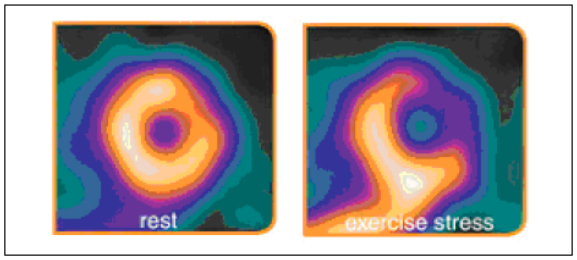


Figure 3. A nuclear image of a heart with CAD at rest and after exercise [16]

V. CBR Techniques for Processing the Myocardial SPECT scintigrams

Recently various CBR approaches such as classification of heart-rate patterns to diagnose stress related disorders were presented, confirming the diagnostic potential for CBR in health sciences and biomedicine [17, 18]. In CBR computing, case library is demonstrated and each case included patient information like sex and age, segmental values of the relative thallium-201 activity obtained by polar map analysis of the scintigraphic images (84 integer values) and 15 integer values representing the results of coronary angiography specifying the location and the severity of stenotic lesions in the 15 segments of coronary arteries.

Aliasghar et al. [17, 18] presents that for identification of similar cases from the case library a difference score was calculated as the weighted sum of differences between the segmental uptakes of each

ISBN: 978-1-61804-236-1

case in the case library with the current index case. The following reasoning techniques are used for determination of presence, location and severity of CAD for myocardial SPECT scintigrams:

A. *GLOB:*

CBR (GLOB) compares all segments of the polar map of the index case with corresponding segments of cases in the library using a calculated similarity measure and retrieves the most similar case (best-match) or a set of similar cases (best-list) to derive the diagnosis for the case.

B. Territorial (TER):

CBR (TER) compares single case which was divided into three separate territories corresponding to the anatomic distribution of the three major coronary arteries and each major vessel territory was treated as an independent case. Each vessel territory (LAD/ LCX/RCA) on the polar map was considered as a separate case, which was compared to corresponding territories of cases in the case library. For calculation of the similarity measure the quantitative polar map was divided into three corresponding territories and retrieves three best-match cases (or three best-lists) for each territory of the LAD, LCX, and RCA to derive the angiographic diagnosis for the corresponding territory.

C. GROUP:

CBR (GROUP) technique divides the case library into partitions with similar cases to optimize the accuracy and efficiency of a system. In this method, the case library was divided in eight case groups corresponding to the extent of coronary artery disease in the three major coronary vessels using the angiography data of each case. Eight case groups were compiled as follows:

i. Group 1: normal cases or cases with only minor lumen
ii. Groups 2 --4: cases with single vessel disease (LAD, LCX, or RCA).
iii. Groups 5 --7: cases with two vessel disease (LAD-LCX, LAD-RCA, or LCX-RCA).
iv. Group 8: cases with three vessel disease (LAD, LCX and RCA).

In this method, the new case was compared with all cases in the eight subgroups separately and a group difference score was calculated as an average of the difference score of cases in the corresponding group. The most similar group (best-match) or a list of groups (best-list) was then used to derive the diagnosis for the presented case.

In the previous CBR techniques, Two retrieval approaches where used:

A. Best-match Approach:

A case (for GLOB and TER method) or a group (for GROUP method) with the highest degree of similarity (lowest difference score) was retrieved from the library as the result.

B. Best-list Approach (Adaptation approach):

A set of cases or groups with the highest degree of similarity (lowest difference score) were retrieved from the case library and were then adapted to resemble more closely the index case.
The results of this study show that the best-match approach of both TER and GROUP retrieval methods showed a higher diagnostic accuracy than the GLOB.

VI. Discussion and Conclusion

Many of the early intelligent reasoning systems were attempted to apply rule-based reasoning approach in developing knowledge-based diagnosis systems in medical domain. However, for a broad and complex medical domain the effort of applying rule-based systems has encountered several problems. Today many systems are serving multi-purpose i.e. tend to support not only in diagnosis but also in number of other complex tasks and combining more than one reasoning techniques in the healthcare domain. These systems can be used to support non experts with a preliminary interpretation in those situations in which experts are not present.

In recent years, machine learning and computational intelligence techniques such as expert systems, genetic algorithms, swarm intelligence, fuzzy logic, neural networks and case-based reasoning have

ISBN: 978-1-61804-236-1

improved the diagnostic accuracy of automated interpretation of myocardial perfusion images [19, 20, 21, and 22]. These methods use the individual diagnostic information of both normal and abnormal images of a case library and thus can potentially offer better diagnostic accuracy than does polar map analysis using a database of reference limits that have been derived by statistical analysis.

However, the medical domain of this study is a suitable and challenging application domain for CBR. And, CBR approach is a robust methodology for determination of presence, location and severity of CAD for myocardial SPECT scintigrams. Clinicians often explain that they reason in terms of similar cases and adapt them to the current situation. A clinician may start his/her practice with some initial past experience (own or learned solved cases), then try to utilize this past experience to solve a new problem and simultaneously increases his/her experience. One main reason that CBR is seen as suitability for the medical domain is its adequate cognitive model and cases may be extracted from the patient's records.

REFERENCES

[1] *Kolodner, J. Case-Based Reasoning, Morgan Kaufmann, San Mateo, (1993).*

[2] *Silvana, Q., Pedro, B., and Steen, A. Proceedings of 8th Conference on Artificial Intelligence in Medicine in Europe, AIME, Cascais, Portogal, Springer, (2001)*

[3] *Hinkle, D. and Toomey, C., Applying Case-Based Reasoning to Manufacturing, AI Magazine, pp. 65-73, (1995)*

[4] *Rissland, E.L. and Danials, J.J., A Hybrid CBR-IR Approach to Legal Information Retrival, Proceedings of the Fifth International Conference on Artificial Intelligence and Law, (ICAIL-95), pp. 52-61, College Park, MD, (1995).*

[5] *Salem, A.M. and Baeshen, N., Artificial Intelligence Methodologies for Developing Decision Aiding Systems, Proceedings of Decision Sciences Institute, 5th International Conference, Integrating Technology and Human Decisions: Global Bridges into the 21st Century (D.I.S. 99 Athens), Greece, pp.168-170, (1999).*

[6] *M. Lenz, S Wess, H Burkhard and B Bartsch, Case based reasoning technology: from foundations to applications, Springer 1998.*

[7] *B Heindl. Et al,: A Case-Based Consiliarius for Therapy Recommendation (ICONS) computer-based advise forv calculated antibiotic therapy in intensive care medicine, computer methods and programs in biomedicine 52, pp 117-127, 1997.*

[8] *Wim van den Broek, Alberto Cuocolo and Adriana Ghilardi, "Myocardial Perfusion Imaging", European Association of Nuclear Medicine Technologist Committee Education Sub-Committee (2004)*

[9] *Kenneth Revett, Abd-badeeh M. Salem and Florin Gorunescu, "A Rule-based Approach to Processing SPECT Imaging for the diagnosis of Heart Disease", In the fifth International Conference of Euro-Mediterranean Medical Informatics and Telemedicine (2009)*

[10] *Aamodt A, Plaza E. Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications 7, pp 39-59, 1994.*

[11] *Bichindaritz I, Marling C. Case-based reasoning in the health sciences: What's next? In Artificial Intelligence in Medicine. 36(2), 2006, pp 127-135*

[12] *Gierl L, Schmidt R. CBR in Medicine. In Case- Based Reasoning Technology, From Foundations to Applications. Springer-verlag. 1998, pp. 273 – 298. ISBN:3-540-64572-1*

[13] *Montani S. Exploring new roles for case-based reasoning in heterogeneous AI systems for medical decision support. In Applied Intelligence. 2007, pp 275–285*

[14 http://www.imagingeconomics.com/issues/articles/MI_2005-06_01.asp

[15] Felipe Massicano, Adriana V. F. Massicano, Natanael Gomes da Silva, Felipe Belonsi Cintra, Rodrigo Müller de Carvalho and Hélio Yoriyaz, "ANALYSIS OF CT AND PET/SPECT IMAGES FOR DOSIMETRY CALCULATION", International Nuclear Atlantic Conference (2009)

[16] http://www.auntminnie.com

[17] Aliasghar Khorsand, Mojgan Haddad, Senta Graf, Deddo Moertl, Heinz Sochor, and Gerold Porenta, "Automated Assessment of Dipyridamole 201Tl Myocardial SPECT Perfusion Scintigraphy by Case-Based Reasoning", The Journal of Nuclear Medicine Vol. 42, No. 2 (2001)

[18] Aliasghar Khorsand , Senta Graf , Heinz Sochor ,Ernst Schuster , Gerold Porenta , "Automated assessment of myocardial SPECT perfusion scintigraphy: A comparison of different approaches of case-based reasoning", Artificial Intelligence in Medicine (2007)

[19] Dan Lindahl, John Palmer, Mattias Ohlsson, Carsten Peterson, Anders Lundin and Lars Edenbrandt, "Automated Interpretation of Myocardial SPECT Perfusion Images Using Artificial Neural Networks", The Journal of Nuclear Medicine Vol. 38, No. 12 (1997)

[20] Ernest V. Garcia, C. David Cooke, Russell D. Folks, Cesar A. Santana, Elzbieta G. Krawczynska, Levien De Braal, and Norberto F. Ezquerra, "Diagnostic Performance of an Expert System for the Interpretation of Myocardial Perfusion SPECT Studies", The Journal of Nuclear Medicine Vol. 42, No. 8 (2001)

[21] Russell D. Folks, "Interpretation and Reporting of Myocardial Perfusion SPECT: A Summary for Technologists", The Journal of Nuclear Medicine, Vol. 30, No. 4 (2002)

[22] Rashid Jalal Qureshi and Syed Afaq Husain, "Design of an Expert System for Diagnosis of Coronary Artery Disease Using Myocardial Perfusion Imaging", National Conference on Emerging Technologies (2004)

ENAMS: Energy Optimization Algorithm for Mobile Sensor Networks

Mohaned. Al. Obaidy
Gulf College, OMAN
Email: mohaned@gulfcollegeoman.com

***Abstract*—This paper presents the design of an intelligent energy optimization algorithm which is based on Swarm Intelligence to increase the life time of swarmed wireless sensor networks. This algorithm represents a further autonomous stage to our previous work which was devoted to cluster Wireless Sensor Networks (WSNs) into independent clusters. Our Algorithm is mainly designed to keep the optimum distribution of clustered sensors while those mobile sensors are directed as a swarm to achieve a given goal. The algorithm presented in this research is suitable for large scale mobile sensor networks and provides a robust and energy- efficient communication mechanism. We are using the Particle Swarm Optimization (PSO) technique to decrease the energy consumption for the entire sensor network. One of the main strengths in the presented algorithm is that the number of clusters within the sensor network is not predefined, this gives more flexibility for the nodes' deployment in the sensor network. Another strength is that sensors' density is not necessary to be uniformly distributed among the clusters, since in some applications constraints, the sensor nodes need to be deployed in different densities depending on the nature of the application.**

***Keywords*—Energy Optimization, Particle Swarm Optimization, Swarm Intelligence, Wireless Sensor Networks**

I. INTRODUCTION

Recent advances in micro-electro-mechanical systems, digital electronics, and wireless communications have led to the emergence of wireless sensor networks (WSNs), which consist of a large number of sensing devices each capable of detecting, processing and transmitting environmental information. A single sensor node may only be equipped with limited computation and communication capabilities; however, nodes in a WSN, when properly configured, can collaboratively perform signal processing tasks to obtain information pertaining to remote and potentially dangerous areas in an untended and robust way. Applications for wireless sensors networks include battlefield surveillance, environmental monitoring, biological detection, smart spaces, industrial diagnostics, etc. [1]. Any WSN is deeply involved in and related to the monitored environment, and any change occurring to the surroundings will significantly influence its performance; nevertheless, the network must be able to tolerate and 'survive'any change by implementing proper reactions and adaptation mechanisms sustaining communications for both sensed data and commands [2].

Energy efficiency has been deemed to be the main challenge for Wireless Sensor Networks. Generally, the power supply of a single sensor node relies on a battery with limited energy (e.g., an AAA battery). Changing or recharging a nodes' battery is very difficult, if not impossible, after sensor nodes have been deployed. Therefore; it is desirable to design energy efficient protocols to run on individual nodes, to ensure that the operation time of the deployed WSN is as long as possible. However, some classical information processing approaches do not consider the energy efficiency issue and require re-examination when applied in resource constrained WSNs. Geographically distributed nodes in a WSN may have different views of the physical phenomenon in the sensor field and thus their measurements may have some points of correlation. A well designed algorithm should also exploit this to accomplish the information processing task via collaboration between nodes. In this work we propose to design an algorithm for a large scale mobile sensors network. This algorithm should provide a robust and energy-efficient communication mechanism which enables the swarms of sensors to move while keeping optimum distances between the sensor nodes.

The rest of this paper will be structured into the follwing sections; Section 2 describes a background ideas and motivation for our work. In section 3, we are explaining the concepts of PSO technique to enable the clusters to move as Swarms while keeping the optimum distances. In section 4 the implementation of the proposed algorithm is explained by showing some snapshots of the simulation program. Section 5 shows the results discussion as well as comments for the output graphs are presented here including a critical review. Finally in section 6 we concluded our work and its objectives with possible future development and enhancments.

II. BACKGROUND AND MOTIVATION

As the Internet has revolutionized our life by the uncomplicated exchange of various forms of information among a large number of users, Wireless Sensor Networks (WSNs) may, in the near future, be equally significant in providing information regarding physical phenomena of interest; ultimately leading to detection and control, and where relevant enabling us to construct more accurate models of the physical world. WSNs have gained tremendous importance in recent years because of its potential use in a wide variety of applications. This, along with the unique characteristics of these networks, has spurred a significant amount of research for coming with network protocols specifically tailored for sensor networks [1]. Wireless sensor networks are developing quickly and have been widely used in both military and civilian applications such as target tracking, surveillance, and security management. Since a sensor is a small, lightweight, un-tethered, battery-powered device, it has limited energy [3]. Therefore, energy consumption is a critical issue in sensor networks. We are interested in sensor networks in which a large number of sensors are deployed to achieve a given goal. All data obtained by member sensors must be transmitted to a sink or data

ISBN: 978-1-61804-236-1

collector. The longer the communication distance, the more energy will be consumed during transmission [4]. Direct transmission networks are very straightforward to design but can be very power-consuming due to the long distances from sensors to the sink. Alternative designs that shorten or minimize the communication distances can extend network lifetimes. The use of clusters for transmitting data to a base station leverages the advantages of small transmit distances for most nodes, requiring only a few nodes to transmit far distances to the base station. Clustering means to partition the network into a number of independent clusters, each of which has a cluster-head that collects data from all nodes within its cluster [5], [6]. These cluster-heads then compress the data and send it directly to the sink point. Figure 1 shows an example of clustered sensor network.

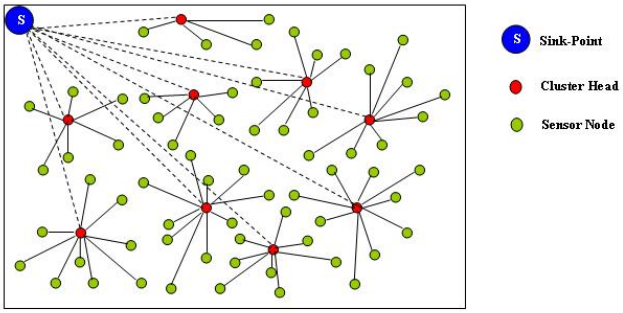


Fig. 1. Clustered Sensors Network

This research represents a further autonomous step to our previous work [7] which was based on Genetic Algorithms (GAs) to divide the WSN into independent clusters. The presented ENAMS algorithm enables clustered WSN to be self organized network while the sensors are moving on a swarm bases. Deployment of mobile swarms can enhance the sensor network in many ways. Firstly, the swarm nodes have much higher hardware capabilities than the sensor nodes. They can provide detailed information of the intended area (e.g. the hot spot). Secondly, the wireless radios of the swarm nodes usually have much longer range and higher channel bandwidth, which can support high quality and delay sensitive multimedia streams. Thirdly, the swarms are mobile [8]. They can be easily directed to the hot spots. A limited number of mobile swarms can easily cover a large scale sensor network. The sensor network can be deployed to cover a very large field due to the low cost of sensor nodes.

A. Energy-Aware Wireless Sensor Networks

Nodes in a WSN are usually highly energy-constrained and expected to operate for long periods from limited on-board energy reserves. To permit this, nodes and the embedded software that they execute must have energy-aware operation. Energy efficiency has been of significant importance since WSNs were first conceived but, as certain applications have emerged and evolved [9], a real need for ultra-miniaturized long-life devices has re-emerged as a dominant requirement. Because of this, continued developments in energy-efficient operation are paramount, requiring major advances to be made in energy hardware, power management circuitry and energy-aware algorithms and protocols.

The energy components of a typical wireless sensor node are shown in Figure 2. Energy is provided to the node from an energy source, whether this is a form of energy harvesting from sources such as solar, vibration or wind, or a resource such as the mains supply or the manual provision and replacement of primary batteries. Energy obtained from the energy source is buffered in an energy store; this is usually a battery or super capacitor. Finally, energy is used by the node's energy consumers; these are hardware components such as; the microcontroller, radio transceiver, sensors and peripherals.

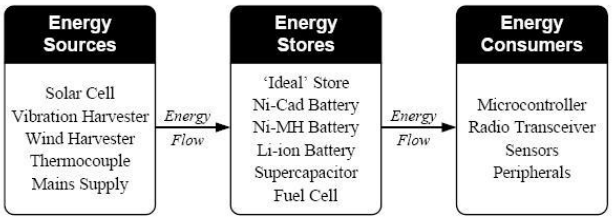


Fig. 2. Energy components of a typical sensor node

With the increased usage of energy sources in nodes [10], [11], the need for energy stores other than batteries (many of which suffer from only offering a limited number of charging cycles) is increased. This can be seen by the researchs that are now utilizing super capacitors (devices that are similar to standard electrolytic capacitors, but with capacities of many Farads) to store the node's energy [11], [12].

To be energy-aware, the embedded software executing on the node must be aware of the state of its energy components. This may be as advanced as monitoring the energy harvested from each source [13], inspecting the rate of consumption by different consumers [14], directing the flow of energy from and to different stores and managing the charging of rechargeable stores [12]. Alternatively, this may equate to simply being able to inspect the residual energy in a single store. Therefore, the embedded software must not only be capable of interfacing with energy hardware (this is generally a requirement of power management circuitry), but also interpreting the data that are obtained usually in the form of a sampled voltage into a remaining lifetime, power or energy. Based upon these values, the operation of the node is adjusted accordingly, usually to maximize the lifetime of the network.

B. Swarm Intelligence

Swarm Intelligence (SI) indicates a recent computational and behavioural metaphor for solving distributed problems that originally took its inspiration from the biological examples provided by social insects (ants, termites, bees, wasps) and by swarming, flocking, herding behaviours in vertebrates [15]. It is an attempt to design algorithms or distributed problem-solving devices inspired by the collective behaviour of social insects and other animal societies. The common behaviours in all kinds of swarms are [15], [16], [17];

- Control is fully distributed among a number of individuals;
- Communications among the individuals happen in a localised way;

ISBN: 978-1-61804-236-1

TABLE I. THE PARAMETERS FOR PSO VELOCITY AND POSITION UPDATE

Parameter	Description
v_i^k	velocity of particle i at iteration k
w	inertia weight
v_i^{k+1}	velocity of particle i at iteration $k+1$
c_j	acceleration coefficients j=1,2
$rand_i$	random number between 0 and 1 i=1,2
s_i^k	current position of particle i at iteration k
$pbest_i$	$pbest$ of particle i
$gbest$	$gbest$ of the group
x_i^{k+1}	position of the particle i at iteration $k+1$

- System-level behaviours appear to transcend the behavioural repertoire of the single individual; and
- The overall response of the system is quite robust and adaptive with respect to changes in the environment.

Swarm intelligence as defined by Bonabeau, Dorigo and Theraulaz is "any attempt to design algorithms or distributed problem-solving devices inspired by the collective behaviour of social insect colonies and other animal societies" [16]. The term "swarm" is used in a general sense to refer to any such loosely structured collection of interacting agents. The classic example of a swarm is a swarm of bees, but the metaphor of a swarm can be extended to other systems with a similar architecture. An ant colony can be thought of as a swarm whose individual agents are ants, a flock of birds is a swarm whose agents are birds, traffic is a swarm of cars, a crowd is a swarm of people, an immune system is a swarm of cells and molecules, and an economy is a swarm of economic agents. Although the notion of a swarm suggests an aspect of collective motion in space, as in the swarm of a flock of birds, all types of collective behaviour are considered here, not just spatial motion.

III. PSO BASED MOVABLE CLUSTERS

Our algorithm is designed to provide the distance management by using Particle Swarm Optimization (PSO) which makes the wireless sensor network self organised while the sensors are moving on a swarm bases. In PSO, the potential solutions are called particles, fly through the problem space by following the current optimum particles. The particles are initialised randomly [18]. Each particle will have a fitness value, which will be evaluated by the fitness function to be optimised in each generation. Each particle knows its best position $pbest$ and the best position so far among the entire group of particles $gbest$. The particle will have velocities, which direct the flying of the particle. In each generation the velocity and the position of the particle will be updated. The velocity and the position update equations are given below as (1) and (2) respectively.

$$v_i^{k+1} = wv_i^k + c_1 rand_1 * (pbest_i - s_i^k) + c_2 rand_2 * (gbest - s_i^k) \quad (1)$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \quad (2)$$

The parameters used in equations 1 and 2 are described in Table I.

The pseudo code for our proposed algorithm is shown in Algorithm (1).

In recent times, there has been a number of improvements to the original PSO [19]. In this paper we have explored two versions of PSO where the extension to the original algorithm is distinct from each other. These are discussed in the following sections.

PSO Initialization: Assume the initial population for PSO is the best solution generated by GAs from previous stage;
while *the stop condition is not satisfied* **do**
  Evaluate the fitness value for each particle's position in the swarm;
  if $fitness(p)$ *better than* $fitness(pbest)$ **then**
    $pbest = p$;
    Set best of $pbest$ as $gbest$;
  end
  Update the particles' velocity v_i^{k+1};
  Update the particles' position x_i^{k+1};
end

Algorithm 1: Pseudo code for PSO Algorithm

A. PSO - Time Varying Inertia Weight (TVIW)

PSO-TVIW model is the same basic PSO algorithm with inertia weight parameter is varying with time from 0.9 to 0.4 and the acceleration coefficient is set to 2. This model is proposed by [20]. The time varying inertia weight is mathematically represented as follows:

$$w = (weight - 0.4) * \frac{(MAXITER - iter)}{MAXITER} + 0.4 \quad (3)$$

Where, MAXITER is the maximum iteration allowed, $iter$ is the current iteration number and $weight$ is a constant set to 0.9.

B. Particle Swarm Optimisation with Supervisor-Student Model (PSO-SSM)

In this method [21] proposed PSO-SSM to achieve low computational costs. The algorithm introduces a new parameter called momentum factor (mc) to update the positions of particles. In this algorithm, they also proposed a different velocity updation mechanism from the conventional PSO algorithms. Here velocity is updated only if each particle's fitness at the current iteration is not better than that of previous iteration. The velocity serves as a navigator (supervisor) by getting the right direction, while the position (student) gets a right step size along the direction. The velocity and the position are modified using the following equations:

$$v_i^{k+1} = v_i^k + c_1 rand_1 * (pbest_i - s_i^k) + c_2 rand_2 * (gbest - s_i^k) \quad (4)$$

$$x_i^{k+1} = (1 - mc) * x_i^k + mc * v_i^{k+1} \quad (5)$$

IV. IMPLEMENTATION AND EXPERIMENTATION

A. Energy Model for Optimisation

We are studying the impact of the transmission range of sensor nodes and positioning of the sink in minimising the communication energy in a sensor network. The important components of each sensor are the data and control processing unit and the radio for communication. The microprocessor used in the processing unit should be energy efficient with

less energy consumption. The energy dissipation in the radio depends on the different characteristics of the radio. The energy model used in this work is adopted from [6], [22], [23] and summarised here. The energy dissipation for transmitting b bits to d distance is shown in Equation 6.

$$E_{tx}(b,d) = E_{elec} \times b + E_{amp} \times b \times d^2 \quad (6)$$

The energy dissipation in a node to receive b bits of data is shown in Equation 7.

$$E_{rx}(b) = E_{elec} \times b \quad (7)$$

Where E_{elec} is the radio energy dissipation and E_{amp} is the transmition amplifier energy energy disipation. Energy consumption of a wireless sensor node transmitting and receiving data from another node at a distance d can be divided into two main components: Energy used to transmit, receive and amplify data and energy used for processing the data, mainly by the microcontroller. Leakage current can be as large as a few mA for the microcontroller, and the effect of leakage current can be neglected for higher frequencies and lower supply voltage. Assuming the leakage current as negligible, the total energy loss for the sensor system due to the distance E_{dd} can be calculated according to Figure 3 using the following equation:

$$E_{dd} = \left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} (d_{ij}^2 + \frac{D_j^2}{n_j}) \right) \quad (8)$$

For more details about the derivation and proof refer to [22].

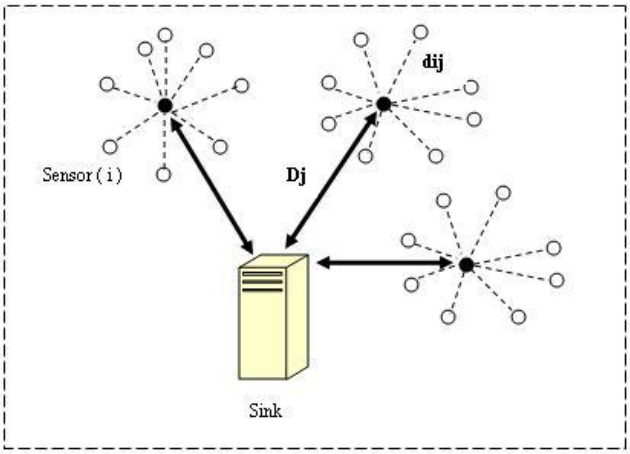


Fig. 3. Energy Model for distance based Sensor Network

B. Experiments and Simulation

In this section, we explore the use of PSO to solve the distance minimization problem for dynamic sensor networks. To implement our algorithm, we have used Java-Applet as a programming environment to simulate the experiments of our algorithm which enables the sensors to move as a swarm using PSO while keeping the optimum distances between the sensor-nodes and their related cluster-head, avoiding any unnecessary movements. Refering to Equation (8), we can conclude that by reducing the distance from a node to the cluster-head and the cluster-head to the sink we can minimise the energy dissipation in a sensor network. In our simulation, we cluster the nodes taking into consideration that each node can transmit or receive data from all the other nodes. Thus, nodes considered in this network do not have transmission range constraint. Sensors are clustered using entirely distance based Equation (8). The fitness function for this method is as follows [24]:

$$Fitness = min \left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} (d_{ij}^2 + \frac{D_j^2}{n_j}) \right) \quad (9)$$

where,

$$\sum_{j=1}^{k} (n_j + k) = N.$$

N is the number of nodes in a network. For our simulations, we used 100-node networks that are uniformly distributed in a 2-Dimensional problem space [0:100,0:100]. We have studied the impact of sink location on the fitness value of the PSO algorithms. In one set of simulations we considered the sink-point to be located at the center of the network (50,50). In another set of simulations we considered the sink-point to be located remotely at (50,180). For both simulations we use the same set of nodes. The maximum number of generations we were running was 1000. The parameters used in the simulations are tabulated in Table II. Snapshots for the mobile swarmed sensor-nodes are shown in Figure 4. Figure 4-a shows the initial distribution for sensor-nodes which is produced by GAs from the previous phase of our algorithm. It can be observed from this distribution that the WSN is clustered into 4-clusters, each one represents a swarm to be directed and controlled by the PSO when it will start running in the second phase of the algorithm. During PSO phase, clusters will be self-organised while they are moving within the experimentation boundaries. This will avoid the mobile sensors to make any unnecessary movements to reserve energy and enlarge the lifetime for each sensor. It is clear from the screen shots shown in Figure 4 - b, c, d, e and f respectively, that the mobile sensors in each cluster keep adjusting their positions during the movements to keep the distances between the sensor-nodes as much as possible the same as it was in the initial distribution.

TABLE II. INITIALISATION AND PARAMETERS RANGE

Parameter	Range
Population size	100
$MAXITER$	1000
v_{max}	100
x_{max}	100
v range	0-100
x range	0-100

V. CRITICAL REVIEW AND RESULTS

In this work we observed the performance in terms of quality of the average optimum value for 10 trials to the **PSO-SSM** and **PSO-TVIW** models which are described earlier. We chose these two methods for the following reasones; the **PSO-SSM** model is the only model which has the ability to stop particles from moving beyond the boundary of the problem space, that is under the influence of mc parameter in it. The **PSO-TVIW** model is almost similar to the basic PSO algorithm with just the inertia weight varying with time

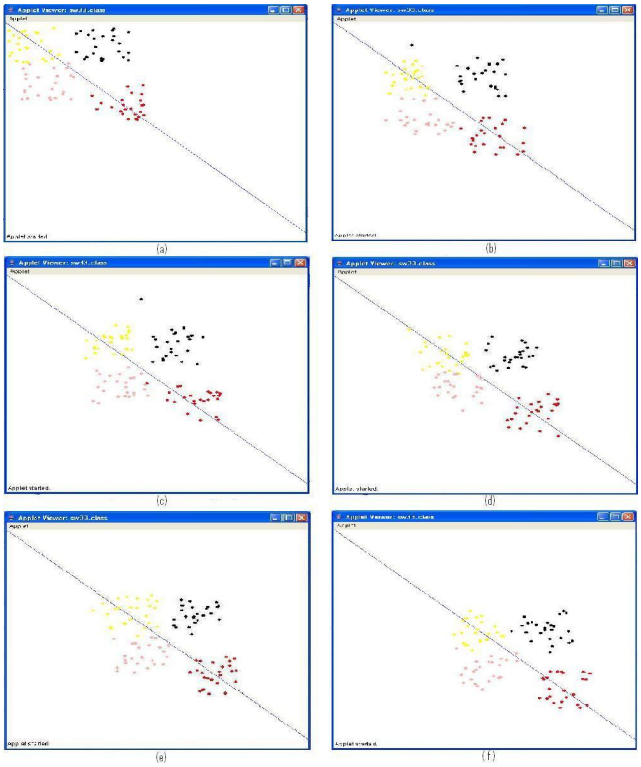


Fig. 4. Snapshots of swarmed WSN with 4 clusters crossing the problem space

from 0.9 to 0.4. From the graph shown in Figure 5 we can conclude that **PS-TVIW** convergence is slower as compared to the **PSO-SSM** algorithm. This was due to constant acceleration co-efficients used in this model which affects the rate of convergence.

Simulation results show that the proposed approach is an efficient and effective method for solving this problem with respect to distance minimization.

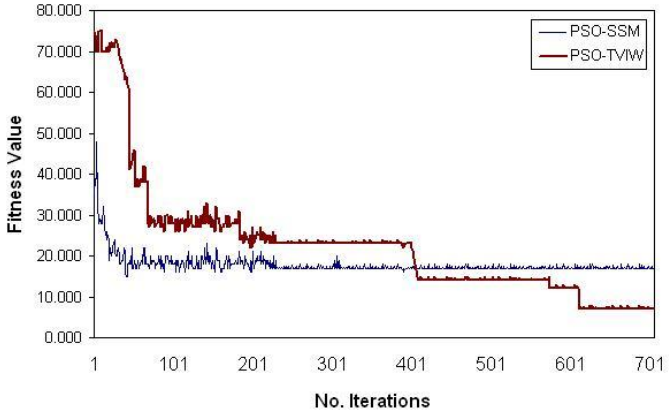


Fig. 5. Convergence for the PSO-SSM and PSO-TVIW Models

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we propose the use of PSO to make WSNs moves as a Swarm keeping the optimum distances between the sensors while they are directed to achieve a specific target. Our proposed approach starts by assuming the initial population for PSO to be the best solution generated by a previous stage of the algorithm which is achieved by using GAs. We also explored the results of the performance evaluation of four extensions to the standard Particle Swarm Optimization algorithm in order to reduce the energy consumption in Wireless Sensor Networks. Communication distance is an important factor to be reduced in sensor networks. We have simulated two models; the Supervisor-Student Model (PSO-SSM) and the time varying Inertia Weight (PSO-TVIW) model. In the (PSO-SSM) model the new parameter introduced called the momentum factor mc to update the position of particles. Also here the velocity is updated only if each particle's fitness at the current iteration is not better than that of previous iteration. Hence the computational costs for this algorithm will be decreased. An important modification proposed is to use boundary checking for re-initialization of particle which moves outside the set boundary. We can also conclude that (PSO-TVIW) convergence is slower as compared to other algorithm. As a future work, our program can be upgraded to cover the two other models described in this paper, then a comprehensive comparison could be done to analyze the behavior of the particles within each case.

We plan to extend the problem on hand by considering a hierarchical structure where a cluster-head can have a super cluster-head which sends data directly to the sink.

REFERENCES

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 8, no. 40, pp. 102–114, August 2002.

[2] S. Bandyopadhyay and E. J. Coyle, "An energy efficient hierarchical clustering algorithm for wireless sensor networks," *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, vol. 3, pp. 1713–1723, 2003.

[3] C. Laurent, D. Helal, L. Verbaere, A. Wellig, and J. Zory, "Wireless sensor networks devices: Overview, issues, state of the art and promising technologies," *ST Journal of Research*, vol. 4, no. 1, pp. 8–11, June 2007.

[4] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy efficient communication protocol for wireless micro-sensor networks," *In Proceedings of the Hawaii International Conference on System Science, Maui, Hawaii*, pp. 3005–3014, 2000.

[5] M. Gerla, T. Kwon, and G. Pei., "On-demand routing in large ad hoc wireless networks with passive clustering," in *In Wireless Communications and Networking Conference (WCNC)*. Chicago, IL, USA: IEEE Computer Society, 2000, pp. 100–105.

[6] W. Heinzelman and A. Chandrakasan, "An application-specific protocol architecture for wireless micro-sensor network," *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 660–670, 2002.

[7] M. Obaidy, A. Ayesh, and A. Sheta, "Optimizing the communication distance of an ad hoc wireless sensor networks by genetic algorithms," *Artificial Intelligence Review, Springer*, vol. 29, no. 3, pp. 183–194, November 2009.

[8] M. Gerla and K. Xu, "Multimedia streaming in larg-scale sensor networks with mobile swarms," *SIGMOD Record*, vol. 32, no. 4, pp. 72–76, December, 2003.

[9] Y. Jin, L. Wang, Y. Kim, and X. Yang, "Eemc: An energy-efficient multi-level clustering algorithm for large-scale wireless sensor networks," *Computer Networks*, vol. 3, no. 53, pp. 542–562, 2008.

[10] C. Park and P. H. Chou, "Ambimax: Autonomous energy harvesting platform for multi-supply wireless sensor nodes," in *Sensor and Ad Hoc Communications and Networks (SECON'06)*, 2006, pp. 168–177.

[11] R. Torah, P. Glynne-Jones, M. Tudor, and S. Beeby, "Energy aware wireless micro-system powered by vibration energy harvesting," in *PowerMEMS*, Freiburg, Germany, 2007, pp. 323–326.

[12] X. Jiang, J. Polastre, and D. Culler, "Perpetual environmentally powered sensor networks," in *4th Int'l Conf. Information Processing in Sensor Networks (IPSN'05)*, Los Angeles, CA, 2005.

[13] A. Weddell, N. Harris, and N. White, "An efficient indoor photovoltaic power harvesting system for energy-aware wireless sensor nodes," in *Eurosensors'08*, Dresden, Germany, 2008, pp. 1544–1547.

[14] T. Stathopoulos, D. McLntire, and W. J. Kaiser, "The energy endoscope: Real-time detailed energy accounting for wireless sensor nodes," in *Int'l Conf. Information Processing in Sensor Networks (IPSN'08)*, St. Louis, MO, 2008, pp. 383–394.

[15] J. Kennedy and R. C. Eberhart, *Swarm Intelligence*. Morgan Kaufman Publishers, 2001.

[16] E. Bonabeau, M. Dorigo, and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*. NY: Oxford Univ. Press, 1999.

[17] M. Dorigo and T. Sttuzle, *Ant Colony Optimization*. MIT press, 2004.

[18] C. Bertelle, M. Obaidy, A. Ayesh, and R. Ghnemat, "Intelligent land-use management and sustainable development: From interacting wireless sensors networks to spatial emergence for decision making," *Engineering of Autonomic and Autonomous Systems, IEEE International Conference and Workshops, Oxford, England*, vol. 0, pp. 73–78, 2010.

[19] M. Obaidy and A. Ayesh, "Energy efficient pso-based algorithm for optimizing autonomous wireless sensor network," in *European Simulation and Modelling (ESM'2008) Conference. EUROSIS, Le Havre, France*, October 2008.

[20] Y. Shi and R. Eberhart, "Empirical study of particle swarm optimization," in *Proceedings of the Congress on Evolutionary Computation, CEC 99*, vol. 3, Washington, DC, USA, July 1999, pp. 1945–1950.

[21] Y. Liu, Z. Qin, and X. He, "Supervisor-student model in particle swarm optimization," in *Proceedings of CEC2004 Congress on Evolutionary Computation*, vol. 1, USA, 2004, pp. 542–547.

[22] S. Guru, A. Hsu, S. Halgamuge, and S. Fernando, "An extended growing self-organising map for selection of clustering in sensor networks," *International Journal of Distributed Sensor Networks*, vol. 1, no. 2, pp. 227–243, 2005.

[23] A. Wang and A. Chandrakasan, "Energy-efficient dsps for wireless sensor networks," *Signal Processing Magazine, IEEE*, vol. 19, no. 4, pp. 68–78, 2002.

[24] M. Obaidy and A. Ayesh, "Optimizing autonomous mobile sensors network using pso algorithms," in *Proceedings of the International Conference on Computer Engineering & Systems (ICCES'08), Egypt*, November 2008, pp. 199–203.

Identification of Direct and Indirect Discrimination in Data Mining

P. Priya and Dr. J. C. Miraclin Joyce Pamila

Department of Computer Science and Engineering, Govt. College of Technology, Coimbatore-13

Abstract--Discrimination is the prejudicial treatment which involves denying opportunities to members of one group in favor of other groups. It is unfair to discriminate people because of their gender, religion, nationality, age and so on, especially when those attributes are used for making decisions about them like giving them a job, loan, insurance, etc. If the training data are inherently biased for or against a particular community, discriminatory decisions may ensue. Discovering the potential biases and eliminating them from the training data without harming their decision-making utility is therefore highly desirable which forms the primary goal of anti-discrimination techniques in data mining. Discrimination can be either direct or indirect. Direct discrimination occurs when decisions are made based on sensitive attributes. Indirect discrimination occurs when decisions are made based on non-sensitive attributes which are strongly correlated with biased sensitive ones. This paper aims at identifying potential discrimination using an acceptable level of discrimination.

Key words--discrimination, direct and indirect discrimination

I Introduction

Discrimination is a very important issue when considering the legal and ethical aspects of data mining. It is more than obvious that most people do not want to be discriminated because of their gender, religion, nationality, age and so on, especially when those attributes are used for making decisions about them like giving them a job, loan, insurance, etc.

It involves denying to members of one group opportunities that are available to other groups. There is a list of **antidiscrimination** acts, which are laws designed to *prevent* discrimination on the basis of a number of attributes (**e.g., race, religion, gender, nationality, disability, marital status, and age**) in various settings (**e.g., employment and training, access to public services, credit and insurance, etc.**). At first sight, automating decisions may give a sense of fairness: classification rules do not guide themselves by personal preferences. However, at a closer look, one realizes that classification rules are actually learned by the system (**e.g., loan granting**) from the training data. If the training data are inherently biased for or against a particular community (**e.g., foreigners**), the learned model may show a discriminatory prejudiced behavior?. In other words, the system may infer that just being foreign is a legitimate reason for loan denial. Discovering such potential biases and eliminating them from the training data without harming their decision making utility is therefore highly desirable. One must prevent data mining from becoming itself a source of discrimination, due to data mining tasks generating discriminatory models from biased data sets as part of the automated decision making. In [12], it is demonstrated that data mining can be both a source of discrimination and a means for discovering discrimination. Discrimination can be either direct or indirect (also called systematic).

1.1 Direct discrimination

Direct discrimination consists of rules or procedures that explicitly mention minority or disadvantaged groups based on sensitive discriminatory attributes related to group membership. Discriminatory (sensitive) attributes like gender, race, religion, etc.,

1.2 Indirect discrimination

Indirect discrimination consists of rules or procedures that, while not explicitly mentioning discriminatory attributes, intentionally or unintentionally could generate discriminatory decisions. Redlining by financial institutions (refusing to grant mortgages or insurances in urban areas they consider as deteriorating) is an archetypal example of indirect discrimination, although certainly not the only one. With a slight abuse of language for the sake of compactness, in this paper indirect discrimination will also be referred to as redlining

ISBN: 978-1-61804-236-1

and rules causing **indirect discrimination will be called redlining rules** [12].

In direct discrimination could happen because of the availability of some background knowledge (rules), for example, that a certain zip code corresponds to a deteriorating area or an area with mostly black population. The background knowledge might be accessible from publicly available data (e.g., census data) or might be obtained from the original data set itself because of the existence of nondiscriminatory attributes that are highly correlated with the sensitive ones in the original data set. .

1.3 Basic definition

Some basic definitions related to data mining [17]. After that, we elaborate on measuring and discovering discrimination.

- **A data set** is a collection of data (records) and their attributes. Let DB be the original data set.
- *An* **item** is an attribute along with its value, e.g., Race = black.
- *An* **item set** is a collection of one or more items, e.g.,{ Foreign worker = Yes, City = NYC}.
- *A* **classification rule** is an expression X -> C, where C is a class item (a yes/no decision), and X is an item set containing no class item, e.g., {**Foreign worker = Yes, City = NYC-> Hire = no}**. X is called the premise of the rule.
- The support of an item set, supp(X), is the fraction of records that contain the item set X. We say that a rule X -> C is completely supported by a record if both X and C appear in the record.
- The confidence of a classification rule, conf(X)->C, measures how often the class item C appears in records that contain X. Hence, if supp(X) > 0 then
- **Conf(X)->C =** $\frac{supp(X,C)}{supp(X)}$

Support and confidence range over (0,1)

- **A frequent classification rule** is a classification rule with support and confidence greater than respective specified lower bounds. Support is a measure of statistical significance, whereas confidence is a measure of the strength of the rule. Let FR be the database of frequent classification rules extracted from DB.
- **Discriminatory attributes and item sets** *(protected by law):* Attributes are classified as discriminatory according to the applicable anti-discrimination acts (laws). For instance, U.S. federal laws prohibit discrimination on the basis of the following attributes: race, color, religion, nationality, sex, marital status, age and pregnancy (Pedreschi et al. 2008). Hence these attributes are regarded as discriminatory and the item sets corresponding to them are called discriminatory item sets. {Gender=Female, Race=Black} is just an example of a discriminatory item set. Let *DAs* be the set of predetermined discriminatory attributes in *DB* and *DIs* be the set of predetermined discriminatory item sets in *DB.*
- **Non-discriminatory attributes and item sets:** If A*s* is the set of all the attributes in *DB* and *Is* the set of all the item sets in *DB,* then *nDA*s (*i.e.* set of *nondiscriminatory attributes*) is *As - DAs* and *nDIs* (*i.e.* set of *non-discriminatory item sets*) is *Is - DIs*. An example of non-discriminatory item set could be {**Zip= 10451, City=NYC**}.

II RELATED WORK

Some proposals are oriented to the discovery and measure of discrimination. The discovery of discriminatory decisions was first proposed by Pedreschi et al. [12], [15]. The approach is based on mining classification rules (the inductive part) and reasoning on them (the deductive part) on the basis of quantitative measures of discrimination that formalize legal definitions of discrimination. For instance, the US Equal Pay Act [18] states that: "a selection rate for any race, sex, or ethnic group which is less than four-fifths of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact."

This approach has been extended to encompass statistical significance of the extracted patterns of discrimination in [13] and to reason about affirmative action and favoritism [14]. Moreover it has been implemented as an Oracle-based tool in [16]. Current discrimination discovery methods consider each rule individually for measuring discrimination without considering other rules or the relation between them. Three approaches are conceivable: **pre-processing, in processing and post-processing approaches**. We next describe these groups

2.1 Pre processing.

Transform the source data in such a way that the discriminatory biases contained in the original data are removed so that no unfair decision rule can be mined from the transformed data and apply any of the standard data mining algorithms. The preprocessing approaches of data transformation and hierarchy-based generalization can be adapted from the privacy

ISBN: 978-1-61804-236-1

preservation literature. Along this line, [7], [8] perform a controlled distortion of the training data from which a classifier is learned by making minimally intrusive modifications leading to an unbiased data set. The preprocessing approach is useful for applications in which a data set should be published and/or in which data mining needs to be performed also by external parties (and not just by the data holder).

2.2 In-processing

Change the data mining algorithms in such a way that the resulting models do not contain unfair decision rules. For example, an alternative approach to cleaning the discrimination from the original data set is proposed in [2] whereby the Non discriminatory constraint is embedded into a decision tree learner by changing its splitting criterion and pruning strategy through a novel leaf relabeling approach. However, it is obvious that in processing discrimination prevention methods rely on new special-purpose data mining algorithms; standard data mining algorithms cannot be used.

2.3 Post processing.

Modify the resulting data mining models, instead of cleaning the original data set or changing the data mining algorithms. For example, in [13], a confidence-altering approach is proposed for classification rules inferred by the CPAR algorithm. The post processing approach does not allow the data set to be published, only the modified data models can be published (knowledge publishing), hence data mining can be performed by the data holder only. One might think of a straightforward pre processing approach consisting of just removing the discriminatory attributes from the data set. Although this would solve the direct discrimination problem, it would cause much information loss and in general it would not solve indirect discrimination. As stated in [12] there may be other attributes (e.g., Zip) that are highly correlated with the sensitive ones (e.g., Race) and allow inferring discriminatory rules.

Preprocessing approach seems to be the most flexible one, it does not require changing the standard data mining algorithms, unlike the inprocessing approach, and it allows data publishing (rather than just knowledge publishing), unlike the post processing approach.

III SYSTEM MODEL

There are two types of rules:
1. PD Rule
2. PND Rule

3.1 Potentially discriminatory rule

A classification rule X -> C is potentially discriminatory (PD) when X = A, B with A is a discriminatory item set and B a nondiscriminatory item set

For example,(Foreign worker = Yes, City = NYC-> Hire = No). The word "potentially means that a PD rule could probably lead to discriminatory decisions. Therefore, some measures are needed to quantify the direct discrimination potential.

3.2 Direct discrimination measure

One of these measures is the extended lift (elift) Let A, B -> C be a classification rule such that Conf (B -> C) > 0. The extended lift of the rule is elift **(A, B -> C)** = $\frac{conf(A,B \rightarrow C)}{conf(B \rightarrow C)}$

The idea here is to evaluate the discrimination of a rule as the gain of confidence due to the presence of the discriminatory items (i.e., A) in the premise of the rule. Whether the rule is to be considered discriminatory can be assessed by thres holding elift as follows.

Let α € R be a fixed threshold and let A be a discriminatory item set. A PD classification rule

c = A, B ->C c is taken as α -discriminatory w.r.t. elift if elift(c) > α. Otherwise is taken as α-protective,. The purpose of direct discrimination discovery is to identify α -discriminatory rules. In fact, α –discriminatory rules indicate biased rules that are directly inferred from discriminatory items (e.g., Foreign worker = Yes).

We call these rules direct α -discriminatory rules. In addition to elift, two other measures slift and olift were proposed by Pedreschi et al. in [13].

ISBN: 978-1-61804-236-1

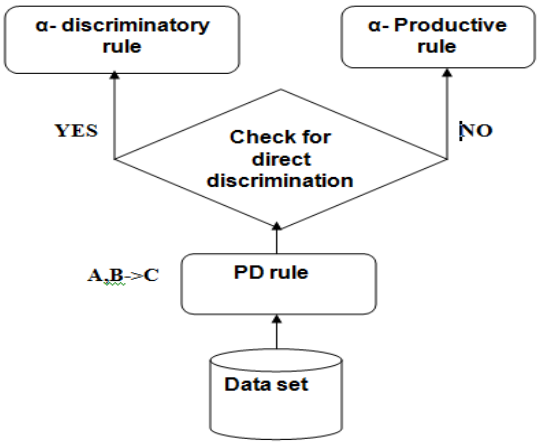


Fig:3.1 Direct discrimination measure

Fig 3.1 says that PD rule of the form A,B->C A is a discriminatory item set and check for direct discrimination such a measure is called Elift .if Elift>α α-discriminatory rule otherwise α-protective rule

3.3 Potentially non discriminatory rule

A classification rule X -> C is potentially nondiscriminatory (PND) when X = D, B is a nondiscriminatory item set. For example,
{Zip =10451, City = NYC -> Hire = No} or
{Experience = Low, City = NYC -> Hire = No}
PND rule could lead to discriminatory decisions in combination with some background knowledge.
e.g., if the premise of the PND rule contains the zip code as an attribute and one knows that zip code 10451 is mostly inhabited by foreign people. Hence, measures are needed to quantify the indirect discrimination potential as well

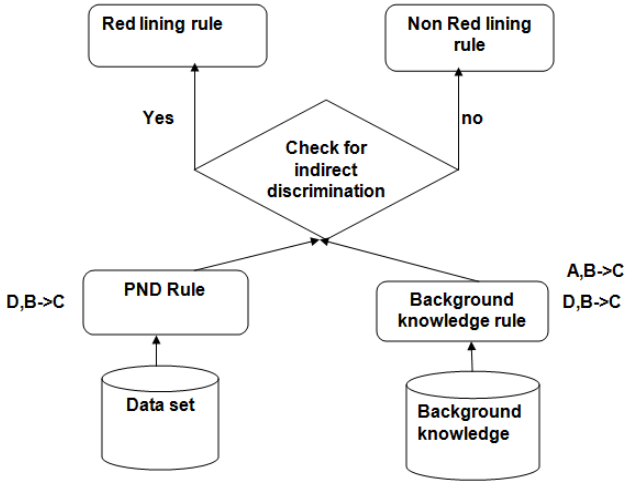


Fig:3.2 Indirect discrimination measure

Fig: 3.2 says that PND rule of the form D,B->C where D is not directly discriminated but highly correlated with discriminatory attribute A . and check for indirect discrimination such a measure is called elb. If elb>α is taken as readlining rule otherwise is taken as non lining rule.

3.5 Indirect discrimination measure

The purpose of indirect discrimination discovery is to identify redlining rules. In fact, redlining rules indicate biased rules that are indirectly inferred from nondiscriminatory items (e.g., Zip = 10451) because of their correlation with discriminatory ones. To determine the redlining rules, Pedreschi et al. in [12] stated the theorem below which gives a lower bound for α discrimination of PD classification rules, given information available in PND rules (γ , δ), and information available from background rules (β1, β2). They assume that background knowledge takes the form of classification rules relating a nondiscriminatory item set D to a discriminatory item set A within the context B.

Theorem Let r: D, B -> C be a PND classification rule, and let
γ= conf(r: D, B -> C) δ= conf (B -> C) > 0:
Let A be a discriminatory item set, and letβ1, β2 such that
Conf (rb1: A, B -> D) ≥β1
Conf (rb2: D, B -> A) ≥β2

$$\mathbf{F(x)} = \frac{\beta 1}{\beta 2}(\beta 2 + x - 1)$$

$$\mathbf{elb(x, y)} = \begin{cases} \frac{f(x)}{y} \text{ if } f(x) > 0 \\ 0 \text{ otherwise} \end{cases}$$

It holds that, for α ≥ 0, if elb (γ, δ) ≥ α, the PD classification rule r': A,B -> C is α-discriminatory. Based on the above theorem, the following formal definitions of redlining and non redlining rules are presented:

A PND classification rule r : D,B -> C is a redlining rule if it could yield an α-discriminatory rule r' : A,B -> C in combination with currently available background knowledge rules of the form rb1 : A,B -> D and rb2 : D,B -> A, where A is a discriminatory item set. For example,
{Zip= 10451, City = NYC}-> Hire = No}.

A PND classification rule r : D,B -> C is a non redlining or legitimate rule if it cannot yield any α-discriminatory rule r': A,B -> C in combination with currently available background knowledge rules of the form

rb1 : A,B -> D and rb2 : D,B -> A, where A is a discriminatory item set. For example,
{Experience = Low, City = NYC} -> Hire= No}.

IV EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Data sets

Two data sets are considered: adult and German credit data set.

4.1.1 Adult data set: We used the Adult data set [10], also known as Census Income, in our experiments. This data set consists of 48,842 records, split into a "train" part with 32,561 records and a "test" part with 16,281 records. The data set has 14 attributes (without class attribute). We used the "train" part in our experiments. The prediction task associated with the Adult data set is to determine whether a person makes more than 50K$ a year based on census and demographic information about people. The data set contains both categorical and numerical attributes. For our experiments with the Adult data set, we set DIs= {Sex = Female, Age = Young}. Although the Age attribute in the Adult data set is numerical, we converted it to categorical by partitioning its domain into two fixed intervals: Age <= 30 is renamed as Young and Age > 30 is renamed as old.

4.1.2 German credit data set: we also used the German Credit data set [11]. This data set consists of 1,000 records and 20 attributes (without class attribute) of bank account holders. This is a well-known real-life data set, containing both numerical and categorical attributes. It has been frequently used in the antidiscrimination literature [12], [7]. The class attribute in the German Credit data set takes values representing good or bad classification of the bank account holders. For our experiments with this data set, we set DIs = {Foreign worker = Yes, Personal Status =Female and not Single, Age = Old}; (cut-off for Age = Old: 50 years old).

4.2 Experimental result for adult data set

```
Rule : martial_status='Never-married' and gender='Male' => salary='<=50K'

A : martial_status='Never-married'
B : gender='Male'
C : salary='<=50K'

Number of tuples which satisfy A , B and C : 5591
Number of tuples which satisfy A and B : 5916
Number of tuples which satisfy B and C : 15128
Number of tuples which satisfy B : 21790

Confidence of A , B -> C : 0.9450642325895876
Confidence of B -> C : 0.6942634235888022
Elift : 1.3612473313145896
Rule : relationship='Own-child' and race='White' => salary='<=50K'

A : relationship='Own-child'
B : race='White'
C : salary='<=50K'

Number of tuples which satisfy A , B and C : 4196
Number of tuples which satisfy A and B : 4255
Number of tuples which satisfy B and C : 20699
Number of tuples which satisfy B : 27816

Confidence of A , B -> C : 0.9861339600470035
Confidence of B -> C : 0.7441400632729365
Elift : 1.3251993928531547
Rule : age='old' and captial_loss='Zero' => salary='<=50K'

A : age='old'
B : captial_loss='Zero'
C : salary='<=50K'

Number of tuples which satisfy A , B and C : 9652
Number of tuples which satisfy A and B : 10275
Number of tuples which satisfy B and C : 23974
Number of tuples which satisfy B : 31041

Confidence of A , B -> C : 0.939367396593674
Confidence of B -> C : 0.7723333655487903
Elift : 1.216271934498383
```

Fig4.1 Direct discrimination measure for adult data set

The α- discriminatory rule identification has been done using the following series of steps:

1. For the given data set, association rules have been generated.

2. From the set of rules, PD rules have been extracted where each PD rule contains at least one discriminatory attribute.

3. For the given data set, elift measure for all the PD rules has been calculated**.**

if elift value is > α,

α discriminatory rule

else

α- protective rule

Table 4.1 gives the number of discriminatory rule for various value of threshold

Table 4.1 Result of Adult data set

Threshold	No of discriminatory rule
0.0	392
0.4	392
0.8	352
0.9	346
1.0	222
1.1	157
1.2	63
1.3	35
1.4	16
1.5	6
1.7	6
1.8	0

From the above table it is seen that as the threshold value is increased, number of discriminatory rule decreases. The maximum threshold value is 1.8 because it gives 0 discriminatory rule. The maximum number of discriminatory rule is 392 because it taken at threshold value 0. Taking an approximately intermediate value for both threshold and number of discriminatory rule the α value has been chosen to be 1and also analyzing set of discriminatory rule for both the previous threshold value 0.9 and next threshold value 1.1seems to be a suitable choice for α

4.3 Experimental result for German credit data set

```
Rule : purpose='A43' and foreign='A201' => class='One'

A : purpose='A43'
B : foreign='A201'
C : class='One'

Number of tuples which satisfy A , B and C : 213
Number of tuples which satisfy A and B : 275
Number of tuples which satisfy B and C : 667
Number of tuples which satisfy B : 963

Confidence of A , B -> C : 0.7745454545454545
Confidence of B -> C : 0.6926272066458983
Elift : 1.1182717732043068
Rule : property='A121' and age='young' => class='One'

A : property='A121'
B : age='young'
C : class='One'

Number of tuples which satisfy A , B and C : 191
Number of tuples which satisfy A and B : 247
Number of tuples which satisfy B and C : 609
Number of tuples which satisfy B : 875

Confidence of A , B -> C : 0.7732793522267206
Confidence of B -> C : 0.696
Elift : 1.1110335520498 86
Rule : property='A121' and foreign='A201' => class='On

A : property='A121'
B : foreign='A201'
C : class='One'

Number of tuples which satisfy A , B and C : 203
Number of tuples which satisfy A and B : 263
Number of tuples which satisfy B and C : 667
Number of tuples which satisfy B : 963

Confidence of A , B -> C : 0.7718631178707225
Confidence of B -> C : 0.6926272066458983
Elift : 1.114399074227145
```

Fig4.2 direct discrimination measure for German credit data set

Table 4.2: Result of German credit data set

Threshold	No of discriminatory rule
0.0	68
0.4	68
0.8	68
0.9	66
1.0	37
1.1	10
1.2	3
1.3	0

From the above table it is seen that as the threshold value is increased, number of discriminatory rule decreases. The maximum threshold value is 1.3 because it gives 0 discriminatory rule. The maximum number of discriminatory rule is 68 because it taken at threshold value 0. Taking an approximately intermediate value for both threshold and number of discriminatory rule the α value has been chosen to be 1and also analyzing set of discriminatory rule for both the previous threshold value 0.9 and next threshold value 1.1 seems to be a suitable choice for α

4.4 Indirect discrimination measure for adult data set

```
Rule : native_country='United-States' and work_class='Private' => salary='<=50K

D : native_country='United-States'
B : work_class='Private'
C : salary='<=50K'
A : race='White'

 Number of tuples which satisfy D , B and C :15594
 Number of tuples which satisfy  D and B : 20135
 Number of tuples which satisfy  B and C : 17733
 Number of tuples which satisfy  B :22696
 Number of tuples which satisfy A , B and D :17728
 Number of tuples which satisfy  A and B : 19404
 Number of tuples which satisfy D , B and A : 17728
 Number of tuples which satisfy  D and B : 20135
Confidence of A , B -> D :0.9136260564831994
Confidence of D , B -> A :0.8804569158182269
Confidence of D , B -> C :0.7744723118947107
Confidence of  B -> C :0.7813271060979908
 Function value : 0.679602143887858
 Elb : 0.8698048980814767
```

Fig4.3 Indirect Discrimination measure for adult data set

Fig 4.3 shows that indirect discrimination measure for Adult data set

ISBN: 978-1-61804-236-1

4.5 Indirect discrimination measure for German credit data set

```
Rule : exp='A73' and no_of_credit='One' => class='One'

D : exp='A73'
B : no_of_credit='One'
C : class='One'
A : age='young'

Number of tuples which satisfy D , B and C : 154
Number of tuples which satisfy D and B : 226
Number of tuples which satisfy B and C : 433
Number of tuples which satisfy B : 633
Number of tuples which satisfy A , B and D : 209
Number of tuples which satisfy A and B : 558
Number of tuples which satisfy D , B and A : 209
Number of tuples which satisfy D , B: 226

Confidence of A , B -> D : 0.37455197132616486
Confidence of D , B -> A : 0.9247787610619469
Confidence of D , B -> C : 0.6814159292035398
Confidence of B -> C : 0.684044233807267
Function value : 0.24551971326164873
Elb : 0.3589237378628722
```

Fig 4.4 shows that indirect discrimination measure for German credit data set

V CONCLUSION

The purpose of this paper is to measure direct and indirect discrimination and identify categories and groups of individuals that have been directly discriminatory in the decision-making processes. The choice of the acceptable level of discrimination has been made. By analyzing the measures of each classification rule, direct discriminatory decision rules have been identified in order to convert them into legitimate (nondiscriminatory) classification rules. The experimental results reported demonstrate that the proposed techniques are quite successful.

REFERENCES

[1] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. Manuscript, 2012.
R. Agrawal and R. Srikant, "Fast Algorithms for MiningAssociation Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases, pp. 487-499, 1994.
[2] T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277-292, 2010.
[3] European Commission, "EU Directive 2004/113/EC on Anti-Discrimination," http://eur-lex.europa.eu/LexUriServ/ LexUriServ.do?uri=OJ:L:2004:373:0037:0043:EN:PDF, 2004.
[4] European Commission, "EU Directive 2006/54/EC on Anti-Discrimination," http://eur-lex.europa.eu/LexUriServ/ LexUriServ.do?uri=OJ:L:2006:204:0023:0036:en:PDF, 2006.
[5] S. Hajian, J. Domingo-Ferrer, and A. Martı´nez-Balleste´, "Discrimination Prevention in Data Mining for Intrusion and Crime Detection," Proc. IEEE Symp. Computational Intelligence in Cyber Security (CICS '11), pp. 47-54, 2011.
[6] S. Hajian, J. Domingo-Ferrer, and A. Martı´nez-Balleste´, "Rule Protection for Indirect Discrimination Prevention in Data Mining," Proc. Eighth Int'l Conf. Modeling Decisions for Artificial Intelligence (MDAI '11), pp. 211-222, 2011.
[7] F. Kamiran and T. Calders, "Classification without Discrimination," Proc. IEEE Second Int'l Conf. Computer, Control and Comm. (IC4 '09), 2009.
[8] F. Kamiran and T. Calders, "Classification with no Discrimination by Preferential Sampling," Proc. 19th Machine Learning Conf. Belgium and The Netherlands, 2010.
[9] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning," Proc. IEEE Int'l Conf. Data Mining (ICDM '10), pp. 869-874, 2010.
[10] R. Kohavi and B. Becker, "UCI Repository of Machine Learning Databases," http://archive.ics.uci.edu/ml/datasets/Adult, 1996.
[11] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "UCI Repository of Machine Learning Databases," http://archive. ics.uci.edu/ml, 1998.
[12] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-Aware Data Mining," Proc. 14th ACM Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), pp. 560-568, 2008.
[13] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring Discrimination in Socially-Sensitive Decision Records," Proc. Ninth SIAM Data Mining Conf. (SDM '09), pp. 581-592, 2009.
[14] D. Pedreschi, S. Ruggieri, and F. Turini, "Integrating Induction and Deduction for Finding Evidence of Discrimination," Proc. 12th ACM Int'l Conf. Artificial Intelligence and Law (ICAIL '09), pp. 157- 166, 2009.
[15] S. Ruggieri, D. Pedreschi, and F. Turini, "Data Mining for Discrimination Discovery," ACM Trans. Knowledge Discovery from Data, vol. 4, no. 2, article 9, 2010.
[16] S. Ruggieri, D. Pedreschi, and F. Turini, "DCUBE: Discrimination Discovery in Databases," Proc. ACM Int'l Conf. Management of Data (SIGMOD '10), pp. 1127-1130, 2010.
[17] P.N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Addison-Wesley, 2006.
[18] United States Congress, US Equal Pay Act, http://archive. eeoc.gov/epa/anniversary/epa-40.html, 1963.
[19] V. Verykios and A. Gkoulalas-Divanis, "A Survey of Association Rule Hiding Methods for Privacy," Privacy-Preserving Data Mining: Models and Algorithms, C.C. Aggarwal and P.S. Yu, eds., Springer, 2008. Sara Hajian

Priya.P received BE degree in Computer Science and Engineering from Jayaraj Anna packiam CSI College of engineering, Nazareth. She is currently doing ME in the Department of Computer Science and Engineering of Government College of Technology, Coimbatore.

Miraclin Joyce Pamila J.C. is an Assistant Professor (Senior Grade) in the Department of Computer Science and Engineering, Government College of Technology, Coimbatore, Tamilnadu, India. She received her Master and Doctoral degree in Computer Science and Engineering from Anna University, Chennai, India. Her fields of interest include Mobile Computing, Data Management Systems, Network Security and Recovery system design for mobile networks. She is a life member of ISTE. She teaches and guides courses at both B.E. and M.E. levels in Computer Science and Information Technology. She has published 30 technical papers in national and international conferences and journals.

Liability for own device and data and applications stored therein

Jan Kolouch, Andrea Kropáčová

***Abstract*—**The number and intensity of cyber attacks and crime online have risen recently, so did their severity and their impact. When an attack takes place (e.g. DoS and DDoS attacks), people start to ask if it is possible the originator of the attack to punish and prosecute (in case he is found). Some people also ask if end-user could be punish and prosecute if his personal computer participates in the attack, for example if his computer is part of botnet. The paper describes the liability for own device and data and application stored therein and reflects on whether end-users should be responsible for their computers. The article describes the legal responsibility of the user in the Czech Republic (different criminal responsibilities are presented on particular articles of the Czech Criminal code), but main principles can be used in any democratic country which prosecute the criminal activities in the Internet.

***Keywords*—**Liability for own computer, botnet, DoS, DDoS, cybercrime, Civil Code, Czech Criminal Code.

I. Introduction

At the beginning of March (4^{th} – 7^{th}) 2013, web services provided in the Czech Republic became the target of a series of (D)DoS attacks [1].

The attacks targeted a different group of web servers every day – the web servers of the most popular news' media on Monday, the oldest and most widely used search engine Seznam.cz on Tuesday, web servers of several Czech banks on Wednesday and web servers of two mobile operators on Thursday. The series of attacks were well prepared – including the planning, selection of targets, good order of the targets, volume of the attacks and used methods.

The Czech Republic and its Internet community gained a lot of valuable experience from this incident and the community and responsible bodies started to talk more about this issue. The discussions were launched at multiple levels and on multiple topics. One of them addressed the issue *"what is the extent of the end user's responsibility when his personal computer was a part of the DDoS attack since his computer was a part of a botnet from which the DDoS attack originated"*, in other words a discussion on *"what is the end user's liability for his own computer and data and applications stored on it, or what is the end user's liability when his computer has been misused for an attack (a crime)"*.

This article addresses the issue of the "liability of the user for own device and data and applications stored therein".

For a better understanding of the article, we first explain what it is the DoS and DDoS attack, remind the role of user in the process of network and services security and basic principles of creating a *botnet*.

II. DoS and DDoS attack

In the area of Internet, a denial-of-service attack (DoS attack) or distributed denial-of-service attack (DDoS attack) is an attack to make a machine, network resource or service unavailable to its users.

The main goal of the attack is basically very simple – to disrupt communication between a user and a service (server) so that the service (server) is unavailable for the user, or at least very slow. The attackers using DoS/DDoS attacks typically target sites or services hosted on high-profile web servers such as banks, credit card payment gateways, and also nameservers.

DoS/DDoS attacks often originate in computers whose owners do not know about the activities of their own computer. This is the result of a successful attack infecting these systems e.g. by malware and their involvement in the botnet network [11].

III. The users

Common users (end-users) do not even realize their role in fighting the cyber attacks and the fact that generally they are an important part of the network and services security. The end user is the key to the security of the network.

Very often, the users' personal computers are attacked, compromised (e.g. by malware) and become a part of a *botnet*. The botnets enable generating massive DoS/DDoS attacks, or hide the activities of the attacker and his identity in executing more sophisticated and precisely targeted attacks with a severe impact.

Unfortunately, the approach of many users to the security of personal computer technology is rather lax, usually filled with the words "I have nothing interesting on my computer (I only play a game), I am not using e.g. the Internet banking, I'm not interesting for anyone, so what ...". This approach could actually constitute a failure to take due care of one's personal computer, the user could lose access to the services, his sensitive personal data may be disclosed, he may incur financial losses or lose access to the Internet. In the worst case

This work has been supported by the CESNET, a. l. e., http://www.cesnet.cz, operator of the Czech national research and education network referred to as CESNET2 within it "Large Infrastructure" (LM2010005) research programme of the Ministry of Education, Youth and Sports of the Czech Republic, running within 2010-2015 timeframe, .

Jan Kolouch works in CESNET, a. l. e., Zikova 4, Prague 6, Czech Republic, (email: kolouch@cesnet.cz).

Andrea Kropáčová works in CESNET, a. l. e., Zikova 4, Prague 6, Czech Republic, (email: andrea@cesnet.cz).

scenario, the user may become an unwilling participant in a crime. The whole Internet may suffer as a result of the irresponsible behaviour of users.

IV. How to Create a Botnet

There are several possibilities how to create a botnet. Yet their fundamental principle is the same – the attacker attempts to control a large number of machines. The machines under control communicate with the controller and receive commands. The infection affects more and more machines.

The result is a botnet which controls tens of thousands of computers, with a powerful computer performance, distributed all over the world, which reduces the means of defence (against "Distributed Denial Of Service") as it is difficult to distinguish the legal traffic from the illegal traffic and implement the appropriate defence.

Let's describe how to create a botnet using Linux systems. Step number one is to obtain access to a machine we want to add to the botnet. One of the methods to obtain access to a Linux machine is to misuse user login information (user name and password) e.g. to the ssh service (the remote access). In general, the attacker advances in following four steps:

1) Gaining access to (most frequently by stealth or interception of) passwords, or keys to access the passwords.
2) Gaining administrator (root) access, in other words, each system is the most vulnerable from within.
3) Installing keyloggers, rootkits etc.
4) Installing malware, a botnet client which is controlled by the controller and performs what it is ordered.

The attacker can apply two basic methods to obtain login information and passwords. The first method is an SSH attack in which user passwords are searched for by force or using a dictionary. The second method is to steal user passwords or keys from previously compromised machines. Step 2 and 3 are optional and depend on the type of attack.

Once the attacker obtains access to a computer, he launches a simple bot-software under the compromised account. Most often, this is some IRC client which can be controlled through IRC channels. Such bots are subsequently used e.g. to SSH attacks and to compromising further accounts.

Another type of attacks is the sophisticated attack in which public or (in exceptional cases) zero-day vulnerability is abused to gain the administrator access and subsequently install malware. Malware is generally hidden so that an ordinary system check cannot detect it. In worst case scenario, this is a rootkit which ensures the running of malware at the core level. In general, malware performs three tasks, yet not all of them need to be implemented.

In any case, the key to including a computer into a botnet is the fact that the machine has been compromised. This is possible because the administrator (owner, user) of the computer concerned:

- Has not configured the system properly
- Does not update the system
- Has not secured it, i.e. does not use firewall, patches, has not installed any antispam (and also malware, spyware) protection
- Has set weak passwords, or is not able to keep the passwords secret
- Has not protected the physical access to computer
- Etc.

Overall, this can be described as a failure to take due care of the computer, which can result in the computer being compromised and incorporated into a botnet and subsequently misused to launch a DDoS attack. This could be applied for instance to make competitor's www services unavailable, etc.

V. Liability for Own Device and Data and Applications Stored Therein

To make all participants of an attacks (e.g. DoS, DDoS) criminally liable is not adequate as the criminal law should be used as *ultima ratio* (the last resort). We have already described how law regulates and defines responsibility for the attacks of DoS and DDoS type in the Czech Republic [2].

So let's dedicate the final section of the article to the debate on and the definition of the possible liability of the users in respect of their own ICT devices, including the data and application stored or running on them.

ICT devices contribute towards the functioning of the today's society. On the other hand, they are the means enabling to commit the heaviest attacks in the virtual world, the consequences of which almost always affect the real world (e.g. a network or information system failure, financial means stolen from the victim's account, etc.).

As mentioned above, the criminal law is an *ultima ratio* means. Besides the criminal or administrative liability, **civil liability** should also be considered when it comes to en attacks (e.g. DoS or DDoS).

Many users of **information** and **communication systems** (in particular **computer systems**) do not realise their possible liability for the abuse of information and communication technology they use.

A **computer system** is a functional unit, consisting of one or more computers and associated software, that uses common storage for all or part of a program and also for all or part of the data necessary for the execution of the programme. A computer system may be a stand-alone system (working independently – e.g. a personal computer) or may consist of several interconnected systems (e.g. a network of computers). The full definition is available [12].

A personal computer (including all connected peripherals), an automated teller machine (ATM), a mobile phone, PDA, a play console (e.g. Sony Playstation, PSP, Wii, Xbox 360) are all examples of computer systems. Computer systems also include televisions that enable running of applications, including the access to the Internet; or automobile systems that ensure similar functions. An example of a rather complex computer system is the Internet.

Information and communication systems are items, and whoever disposes of them should **act so as not to cause unreasonable harm to freedom, life, health or property of another person** [3]. Where **the perpetrator, by a wilful breach of good manners, causes harm he is obliged to compensate it;** where he exercises his rights, the perpetrator is liable to compensate the damage only when his primary goal was to cause harm to another person [4].

Such wording of the civil code clearly stipulates the obligation to duly administer the information and communication systems as well as the duty to prevent harm which could be caused as a result of his activity (including the use of ICT within the Internet environment).

Many common users underestimates the prevention and security of ICT devices at their disposal, either by ignorance or willingly (altogether ignoring the danger and failing to install anti-malware, anti-spam and other similar application, or failing to update the installed programmes and applications, etc.).

Determining the nature of the fault in end user's action is crucial when it comes to civil or criminal liability.

This assertion can be supported by the following three real-life cases.

A. Case 1

The user had an illegal copy of Windows XP SP.2 operational system installed on his personal computer, and did not update it on purpose. The user also knowingly installed programmes allowing third parties to manipulate with the computer without his knowledge.

The goal of the user was to release himself from possible criminal liability for an attack carried out through the computer by a third party (e.g. the computer has purposefully been connected to a botnet network).

In practice, users (participants of the attack/mediator of the attack) base their defence on stressing the fact that they were not the attacker who carried out the attack from a particular computer.

In our point of view, the exemption from liability by claiming that a person is not the direct attacker and did not cause a particular attack by their own action is not possible, or to be more precise, such a claim cannot be accepted as absolute.

From the point of view of the criminal law, at least the provisions on complicity and the principles of the accessory nature of complicity should apply as the action of the person which enable or facilitated the commission of a crime (contributed to the crime for instance by **securing the means or removing the barriers**, enticing the victim to the place of the crime, covering while the offence is committed, advice, reaffirming the determination or a promise) can be subsumed under the provisions on the assisting offender [5]. Securing the means covers granting the access to a computer system or its part to enable the commission of a wilful criminal offence.

Where a high level of direct participation of the user on the illegal action of a third person is proven, such user could potentially be considered an accomplice [6] in the crime. The extent of the awareness about the use of the computer to commit an illegal act would be crucial here, as well as the understanding that the action may violate or threaten the interests protected by the Criminal Code [7].

From the point of view of the civil law, such action by the user could be subsumed under the provision of § 2909 of the Civil Code.

B. Case 2

The user had an illegal copy of the Windows XP SP.2 operation system installed on his personal computer, and did not update it on purpose. He had a number of games and application installed on the computer, and breached the copyright in particular by circumventing or suppressing the protection features. The games and application in question were installed using keygens or cracks[1] which contained malware of other attackers. The user was not aware of the fact that other users use his computer.

This is the most common case of computer abuse without the rightful user being aware of it, although the user, by his illegal action (in particular by breaching the copyright) or by mere lack of knowledge of the computer technology caused that his computer was abused to attack by a third person.

From the point of view of the criminal law, the provisions on complicity and the principles of the accessory nature of complicity cannot be applied here, since the person who enable or facilitated another person the commission of a criminal offence, did not act wilfully, and did not help the key offender.

As regards the fault, the provisions on unwilful negligence could be applied to the action of the user of the infected computer, since the offender was not aware that his action could breach or threaten an interest protected by the Criminal Code, although given the circumstance and personal situation of the offender, he could and should have known [8].

As the Criminal Code fails to define unwilful merits of "Unauthorised access to a computer system and data carrier," it is not possible to apply the institutes of the criminal law on this particular case.

From the point of view of the civil law, such user action could be subsumed under provision of § 2912 (1) of the Civil Code: *"Where the perpetrator fails to act in accordance with the behaviour reasonably expected from a person of average character in the private communication, it is assumed he acted in negligence."*

It should be noted that the person who caused the harm (the perpetrator) is obliged to compensate the damage, irrespective of whether or not he is guilty in cases defined by a special

[1] These are third-party interferences into programmes in order to modify them and enable their easy running (the keygens), paralyse the programme protection which prevents its copying or running under pre-defined conditions (the cracks), and other modifications to the programme with the aim of subsequent use or distribution to third parties.

ISBN: 978-1-61804-236-1

regulation [9].

C. Case 3

The user is taking due care of his personal computer (updating SW equipment) and reasonably protects it (using anti-virus, anti-spam and anti-malware protection and checks). Despite that, the computer was attacked by a third party (e.g. connected to a botnet) and subsequently misused to attack another device.

We believe that from the point of view of the fault, the provisions on unwilful negligence do not apply to the user of the affected computer.

Given the pro-active behaviour of the user, provisions of the section 232 of the Criminal Code on *Causing harm to a record in a computer system and on a data carrier and negligent intervention into computer equipment*, cannot be applied either, since the provision presupposes gross negligence. See s. 16 (2) of the Criminal Code: *A criminal offence is committed through gross negligence if the offender's approach to the due diligence. indicates a clear wantonness of the offender to the interests protected by the Criminal Code.*

From the point of view of the civil law, the user action in this case cannot be subsumed under the above mentioned provision of the § 2912 (1) of the Civil Code, as the user acted as could be expected from him. This needs to be considered in a broader context since once the user realises that his ICT devices have been misused to illegal attack against another person, he is obliged to notify without undue delay the person to whom the harm could be caused of the possible consequences (it is disputable whether it is possible, at the moment of the attack to establish who the person actually is). Having fulfilled the reporting duty, the victim is not entitled to damages of the harm which he could have prevented after he had been notified [10].

All circumstance of the case need to be carefully considered in each particular case, and the obligation to repay damages can only be set by the court.

On the other hand, where a user fails to "look after" of his device (i.e. does not secure it, update it, etc.) which is subsequently misused to a DoS or DDoS attack, it can be expected that the court in the proceedings to claim damages can order the user to partially or fully (e.g. where the capacity of a data centre is misused) compensate the damage caused to the victim by means of the user's device.

VI. Conclusion

Based on the above analysis, we believe that technical and legal professionals should cooperate more closely in the fight against cyber attacks or cybercrime. This would enable to revise the legal regulations of individual member states to sanction unwanted Internet action and to enable CERT/CSIRT teams[2] and law enforcement agencies in particular to fully exploit the means and the limits of the law to repress such illegal action.

[2] CERT = Computer Emergency Response Team, CSIRT = Computer Security Incident Response Team.

References

[1] KROPÁČOVÁ Andrea. *(D)DoS attacks targeted web servers operated in Czech Republic*. 17th International Conference on Computers: Recent Advances in Computer Science, Rhodos, 16 July 2013, ISBN: 978-960-474-311-7, ISSN: 1790-5109

[2] KOLOUCH, Jan. *Criminal liability for DoS and DDoS attacks*. 17th International Conference on Computers: Recent Advances in Computer Science, Rhodos, 16 July 2013, ISBN: 978-960-474-311-7, ISSN: 1790-5109

[3] See Art. 2900 of Act no. 89/2012 Sb., the Civil Code of Czech Republic

[4] See Art. 2909 and subsequent of the Civil Code of Czech Republic

[5] See Art. 24 (1)(c) of the Criminal Code of Czech Republic

[6] See Art. 23 of the Criminal Code of Czech Republic

[7] See Art.15 (1)(b) of the Criminal Code of Czech Republic

[8] See Art. 16 (1)(b) of the Criminal Code of Czech Republic

[9] See Art. 2895 of the Civil Code of Czech Republic

[10] See Art. 2092 of the Civil Code of Czech Republic

[11] Daniel Plohmann, Elmar Gerhards-Padilla, Felix Leder. *Botnets: Measurement, Detection, Disinfection and Defence.* [online]. [28. 3. 2014]. Available at: http://www.enisa.europa.eu/activities/Resilience-and-CIIP/critical-applications/botnets/botnets-measurement-detection-disinfection-and-defence

[12] KOLOUCH, Jan and Petr VOLEVECKÝ. *Criminal protection against cyber crime.* Prague: The Police Academy of the Czech Republic in Prague, 2013. ISBN 978-80-7251-402-1

Authors Index

ISBN: 978-1-61804-236-1

ISBN: 978-1-61804-236-1